DOCUMENT RESUME

ED 155 207 95 TM 007 173

AUTHOR Tatsuoka, Kikumi K.; Tatsuoka, Maurice M.
TITLE Time-Score Analysis in Criterion-Referenced Tests. Final Report.
INSTITUTION Illinois Univ., Urbana.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
BUREAU NO BR-6-0554
PUB DATE Feb 78
GRANT NIE-G-76-0087
NOTE 177p.; Not available in hard copy due to marginal legibility of original tables

EDRS PRICE MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS *Computer Assisted Instruction; Criterion Referenced Tests; Data Analysis; Feasibility Studies; Goodness of Fit; *Item Analysis; Mastery Learning; *Mathematical Models; Matrices; Post Secondary Education; *Reaction Time; Scores; Statistical Analysis; Test Interpretation; *Timed Tests
IDENTIFIERS *Computer Assisted Testing; Estimation; Gamma Coefficient; Test Theory; *Weibull Distributions

ABSTRACT

The family of Weibull distributions was investigated as a model for the distributions of response times for items in computer-based criterion-referenced tests. The fit of these distributions were, with a few exceptions, good to excellent according to the Kolmogorov-Smirnov test. For a few relatively simple items, the two-parameter gamma distribution provided better fits. The three parameters of the Weibull distribution were as follows: the location parameter represents the theoretical minimum time required; the scale parameter is related to the mean; and the shape parameter (c) is related to two kinds of difficulty indices. It also appeared that the shape parameter was related to the degree of familiarity of the item and the degree of engagement or involvement of the test takers. A function related to c was the conditional response rate, which is called the hazard rate in system-reliability literature. The c parameter was found to be sensitive to the conceptual difficulty of items that were equal according to traditional difficulty indices. An index was developed for the efficiency of lessons and this, too, was found to be related to the Weibull parameter. Finally, c was judged to be related to the optimal cutoff point for criterion-referenced tests. (Author/CTM)

FINAL REPORT

Project No. 6-0554
Grant No. NIE-G-76-0087

TIME-SCORE ANALYSIS IN CRITERION-REFERENCED TESTS

Kikumi K. Tatsuoka

Computer-Based Education Research Laboratory

and

Maurice M. Tatsuoka

Department of Educational Psychology

College of Education

University of Illinois at Urbana-Champaign

Urbana, Illinois 61801

February, 1978

U.S. Department of Health, Education, and Welfare

National Institute of Education

FINAL REPORT

Project No. 6-0554
Grant No. NIE-G-76-0087

TIME-SCORE ANALYSIS IN CRITERION-REFERENCED TESTS

Kikumi K. Tatsuoka

Computer-Based Education Research Laboratory

and

Maurice M. Tatsuoka

Department of Educational Psychology

College of Education

University of Illinois at Urbana-Champaign

Urbana, Illinois 61801

February, 1978

U.S. Department of Health, Education, and Welfare

National Institute of Education

ABSTRACT

This study investigated the feasibility of using the family of Weibull distrubutions - a family which is widely used in system-reliability analysis - as a model for the distributions of time scores (response times) of items in criterion-referenced tests, lesson segments and entire lessons that were implemented on the PLATO system. The items were those of a series of matrix algebra tests developed for the dual purpose of using in this study and for testing students in three statistics courses at UIUC both before and after they studied our matrix algebra course. The latter provided the lesson segments (including exercises) while the entire lessons came from the Chanute AFB CBE project and deals with special and general vehicle maintenance training.

The fits of the Weibull distributions to these various observed distributions were, on the whole, very good to excellent as gauged by the Kolmogorov- Smirnov goodness-of-fit test. However, for some items (most of which possessed certain exceptional properties in common) the two-parameter gamma distribution offered better fits. The same held true with even greater force for the exercises occurring in the matrix algebra lessons. Tentative explanations of when and why the gamma was better than the Weibull were advanced, but discovery of definitive reasons must await future research.

We would be the first to concede that we have barely scraped the surface in studying the utility of response time (time scores) along with performance scores for analyzing and evaluating data from criterion-referenced tests, both for the purpose of assessing the quality of the tests themselves and for improved testing of the examinees' abilities.

Nevertheless, we believe that we have at least demonstrated the feasability of this approach and hope to have shown that further research along these lines is warranted. In particular, the Weibull distribution in its two-parameter form (which we used in this study), three-parameter form, or two-component composite form - long used by system analysts but apparently not widely known among educational and psychological researchers - seems to bear further investigation for this purpose.

ACKNOWLEDGEMENTS

We wish to acknowledge the services rendered by the following persons:

Bob Baillie, expert programmer and numerical analyst, who has several publications in the Mathematics Of Computation and other journals, wrote most of the main programs.

Tamar Weaver, a highly experienced former programmer at the Ministry of Transportation of Israel, who wrote several statistical analysis routines and transformation programs.

Kay Tatsuoka, junior in mathematics and computer science, at the Massachusetts Institute Of Technology, who wrote several statistical analysis routines and utility programs.

Patrick-Maritz, a research assistant, who made a start for our being able to utilize the LOGIST program.

Jerry Dyer, Mark Bradley, and John Matheny, who served as junior programmers.

Julie Garrard, for editorial work and clerical help.

Curtis Tatsuoka, 15 years of age, for refining screen displays and carrying out other small programming tasks.

Bob Linn and Charles Lewis for having their statistics-course students use our matrix algebra lessons and tests. Also, Larry Francis, former Director of the Military Training Group at CERL for giving us access to data from the Chanute AFB CBE Project.

TABLE OF CONTENTS

TIME-SCORE ANALYSIS IN CRITERION-REFERENCED TESTS

1. INTRODUCTION

It is well known that one of the major problems encountered in psychometric and statistical analyses of criterion-referenced (or domain-referenced) tests stems from the fact that, because they are designed primarily for mastery testing, their scores tend to be uniformly quite high. The consequent lack of variability of scores leads to embarrassingly low reliability and validity coefficients when these are defined in the traditional way in terms of product-moment correlation coefficients. A number of authors (e.g., Harris, 1972; Huynh, 1976; Livingston, 1972) have proposed various approaches to side-stepping this problem of limited score variability by offering alternative measures of reliability and validity.

One approach that does not appear to have been exploited to date, however, is the seemingly obvious one of considering time scores—i.e., the time it takes examinees to respond to items or entire tests (assumed unspeeded)--in addition to performance scores. That there is no dearth of variability in time scores is evident from casual observation. The main reason time scores have not been utilized despite this fact is probably that their accurate recording can take place only in the context of computer aided instruction and testing, which are fairly recent developments. Another possible reason is that response times have widely been regarded as erratic phenomena not exhibiting any law-like behavior and hence not indicative of the extent of knowledge or mastery of a subject matter. (We are here obviously excluding the use of time measures such as response latencies and time taken to learn lists of nonsense syllables, paired associates, etc. that have long and widely been used under the tightly controlled conditions of psychological experiments. Also, we are aware that one of Rasch's models [1960] involves a time measure, viz., the time required by a pupil to read a passage of a given length. But again, the situation here is a relatively controlled one. Reading a particular passage is a much more circumscribed activity than, say, taking an algebra test in which various abilities are brought to play.)

One of the present authors has been working in the field of computer based instruction (specifically the PLATO system at the University of Illinois) for a number of years, and she has hence been informally exploring the utilization of time scores for a long time. The research described in this report is an outgrowth of this sustained interest in time scores and represents a more systematic exploration of their utility. We wish to emphasize, however, that this study makes no attempt to enhance the psychometric properties of criterion-referenced tests by offering alternative measures of reliability and validity based on time scores. (That must be deferred to some future project.) The objective

of the present study, to repeat, is simply to explore in depth how time scores behave and to reveal whatever regularities and potential usefulness they may possess.

One way to check whether a variable is behaving in a systematic fashion is to examine its statistical distribution, and if it seems to be following some identifiable theoretical distribution, to see if some rationale can be adduced to explain why it might be expected to follow that particular distribution. Of course there are any number of theoretical distributions a stochastic variable may seem to be following, so it would be like looking for a needle in a haystack if there weren't some guides as to what sort of distribution might fill the bill. Since Rasch's (1960) work parenthetically alluded to above had led to a two-parameter gamma distribution for the time taken to read a passage of N words, this distribution was a possible candidate. However, Brée (1975) had analyzed some empirical data on problem-solving time (albeit of quite limited scope) which showed that a two-parameter negative exponential distribution offered a better fit than the two-parameter gamma distribution, thus decreasing the attractiveness of the latter.

We were therefore thinking of carrying out a larger-scale replication of Brée's study comparing the relative goodness-of-fit of the gamma and negative exponential distributions, utilizing a large and increasing data base accessible to us (and in part developed by us) on the PLATO system, when a third family of distributions shown to be useful in modeling certain time-score distributions came to our attention. This was the Weibull (1951) distribution which, we learned, had been (and continues to be) extensively used in the context of system-reliability theory: the study of the probability of failure, within a given time span, of a mechanical or electronic system as a function of the probabilities of failure of individual components of the system. We learned of this distribution through the works of Sato (1973) and Takeya, Sato and Sunouchi (1975) who had pioneered[1] its application to the modeling of the cumulative response curve, i.e., the plot of the percentage of students completing an item within a given length of time, against the latter as abscissa.

The justification suggested (although not explicitly stated) by Sato and his coworkers for diverting a distribution found to be descriptive of fatigue or failure time to so remote a field of application as response time for test items is as follows. The test item (or total test, or instructional unit, depending on the level of analysis) is identified with the system whose reliability is being assessed. The

---

[1]It was subsequently brought to our attention that Bargman (1966) had also utilized the Weibull distribution in a study of growth functions.

student's "attacks" on the item correspond to the shocks or wear and tear to which the system is subjected, and the eventual solution of the item is the failure of the system. Farfetched as such identifications may seem, they are not unreasonable. It is plausible to imagine the student to be intent on "cracking the system" by answering the item correctly. The time he takes in doing so--the response time--corresponds to the "survival time" (or "fatigue life") of the system. The only difference is that, whereas in system-reliability analysis we want the survival time to be as long as possible, in test-response data we want it to be as short as possible--especially in criterion-referenced tests. Thus, the use of the Weibull distribution in time-score analysis has some intuitive appeal.

Another reason that encourages at least examining the Weibull distribution for the purpose at hand is that the two-parameter negative exponential distribution advocated by Brée can be regarded as a special case of the Weibull distribution--a three-parameter family--when one of its parameters is equated to unity. (See next section for mathematical demonstration.) When it is recalled that Brée's data base was quite limited--comprising solving-time data from three problems originally fitted to gamma distributions by Restle and Davis (1962) plus those for a fourth problem taken from another source--it is not inconceivable that these sets of data happened to be well modeled by this special case of the Weibull distribution. If so, the psychological arguments invoked by Brée to provide a rationale for the two-parameter negative exponential distribution may hold also for the Weibull distribution.

Thus the thrust of our contemplated study shifted from a gamma vs. negative-exponential comparison to a more general one of investigating the usefulness of the Weibull distribution as a model for time-score data from CR tests in the context of CAI. What is reported in the sequel, therefore, includes but is not confined to a comparison of the gamma and Weibull distributions. It also includes attempts to relate the three parameters of the latter distribution to various psychometrically meaningful indices associated with CR tests and their constituent items, such as difficulty level, ability to differentiate between masters and nonmasters, and so forth.

2. THE WEIBULL DISTRIBUTION: RATIONALE AND DERIVATION

Although an intuitive rationale for the applicability of the Weibull distribution for item (or test) response time was given in the introduction by identifying the solution of an item by a student with the failure of a system in system-reliability theory, this rationale does not lead to a derivation of the distribution (or density) function. In other words, the rationale stated earlier is far from being a set of axioms or postulates from which the mathematical form of the density function logically flows. In the final analysis, as Weibull himself (1951)

and subsequent expositors (e.g., Mann, Schafer and Singpurwalla, 1974) have said, the distribution was empirically discovered rather than axiomatico-deductively derived in the first place. Nevertheless, if something even remotely resembling a postulate (or set of postulates) can be found that makes intuitive sense and at the same time logically implies the mathematical expression for the distribution function, this would lend greatly to the credibility of the distribution. Such a basis has been postulated (albeit as an ex post facto rationalization) by system-reliability researchers in terms of the concept of hazard rate, which is essentially the conditional probability that a system which has survived through time t will fail during an infinitesimal time interval immediately thereafter. Translated to fit the context of item response time, this may be dubbed the conditional response rate and is defined in the following subsection.

2.1 Derivation Based on the Conditional Response Rate

Let us denote by $f(t)$ the probability density that a person randomly selected from the population will respond to a given test item (or any other unit of a test) during the infinitesimal time interval $[t, t + dt]$. (The actual probability that the person will respond to the item in this time interval is $f(t)dt$.) Then the proportion of individuals who will have responded to the item by time t is

$$F(t) = \int_0^t f(u)du,$$

which is the (cumulative) distribution function. It follows that the proportion of individuals who have not responded to the item by time t is $1 - F(t)$. Consequently, the conditional probability that a person will respond to the item during the interval $[t, t + dt]$ given that he or she has not responded to the item up to time t is, by the definition of a conditional probability, given by

$$p(\text{responds in interval } [t,t+dt] \mid \text{has not responded by time } t) = \frac{f(t)dt}{1-F(t)}$$

(From the definition of conditional probability, one might expect to find in the numerator the probability of the joint event "has not responded by time t and responds in interval $[t,t+dt]$." However, a little reflection shows that the simple event "responds in interval $[t,t+dt]$" automatically implies "has not responded by time t." Hence the former simple event is synonymous with the joint event cited, and their

probabilities are identical.) The conditional response rate (CRR) is defined by the expression above exclusive of the differential element dt, and we symbolize it by h(t), keeping the notation commonly used for hazard rate in system-reliability theory. Thus

$$(2.1) \qquad h(t) = \frac{f(t)}{1 - F(t)} .$$

From the concept of hazard rate in general, the corresponding distribution and density functions may easily be derived by elementary calculus, as follows. "Tacking on" the differential element dt in both sides of equation (2.1), replacing f(t)dt by the differential element dF(t) of F(t), and further writing u in place of t (in anticipation of using t for the upper limit of a definite integral), we obtain

$$h(u)du = \frac{dF(u)}{1 - F(u)} .$$

Integrating both sides from a lower limit $u = t_o$ to a general upper limit t, we get

$$\int_{t_o}^{t} h(u)du = -\ell n[1 - F(u)]\Big|_{u=t_o}^{u=t}$$

$$= \ell n[1 - F(t_o)] - \ell n[1 - F(t)]$$

$$= -\ell n[1 - F(t)],$$

if we let t_o be the lower limit of the range of t so that $F(t_o) = 0$. It then follows that

$$1 - F(t) = \exp\left[-\int_{t_o}^{t} h(u)du\right],$$

or

$$(2.2) \qquad F(t) = 1 - \exp\left[-\int_{t_o}^{t} h(u)du\right].$$

Taking derivatives of both sides, we get

$$(2.3) \qquad f(t) = h(t) \exp\left[-\int_{t_o}^{t} h(u)du\right].$$

The last two equations express the distribution function and the density function, respectively, as functions of the CRR h(t) in general. Substituting particular expressions for h(t) in these equations gives rise to particular distribution and density functions. The Weibull distribution results essentially when it is assumed that h(t) is a monotonically increasing function of t, is independent of t, or is a monotonically decreasing function of time. (That is, we forbid h(t) from being a function that first increases with t, reaches a maximum, and then decreases with t, or the other way around. Of course, more complicated behaviors are also forbidden.) Actually, we need to be slightly more specific than merely requiring h(t) to be a monotonic function of t; we must require it to be a monotonic power function of t (like t^m). We further write the expression in a more elaborate form in order to have a "neat" expression for the resulting probability density and distribution functions. Specifically, we postulate that

$$(2.4) \qquad h(t) = \frac{c}{\mu_o^c} (t-t_o)^{c-1}$$

Although this expression looks highly contrived, the multiplier c/μ_o^c may, at this point, be regarded simply as a proportionality constant, and the subtraction of t_o from t merely reflects the fact that t_o is the effective "zero point" on the t scale, for no value of t smaller than this can exist, by the definition of t_o given above. Thus, the expression is no more than a "plain" power function t^m with a shift in origin and a rescaling factor.

It is evident from expression (2.4) that h(t) is an increasing function of t, a constant, or decreasing function of t, according as $c > 1$, $c = 1$, or $c < 1$, respectively, as illustrated in Figure 1. From the meaning of h(t), the intuitive (although somewhat loose) interpretations of the three cases are as follows:

1. When $c > 1$, the longer a person persists with the item without responding to it, the more likely it becomes that he/she will answer it "the next moment" (which is roughly what the interval $[t, t+dt]$ means);

2. When $c = 1$, the chances that a person will respond the next moment, when he/she hasn't responded so far, neither increase nor decrease with time;

3. When $c < 1$, the longer a person persists with the item without responding to it, the less likely it becomes that he/she will answer it the next moment.

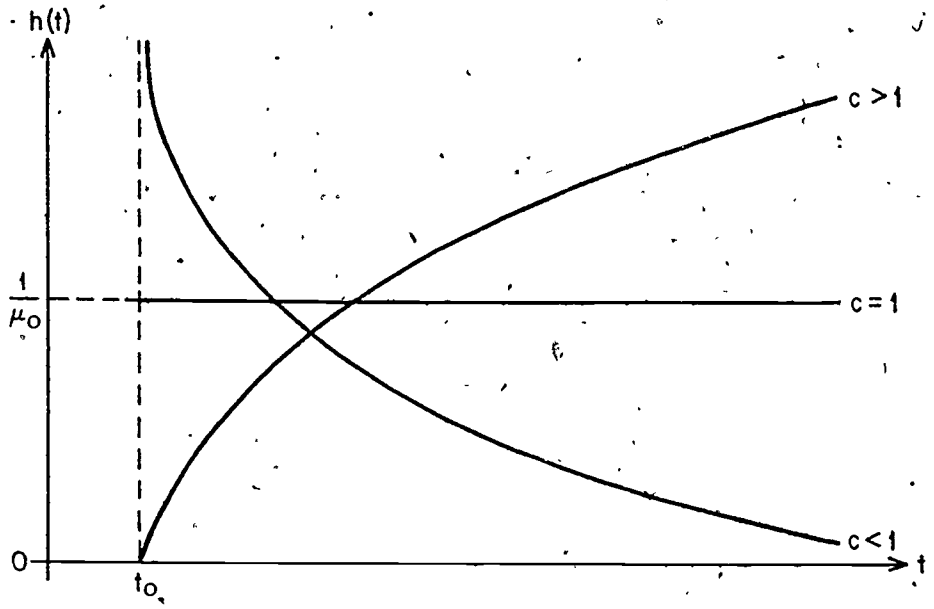


Figure 1. The conditional response rate (CRR) h(t) for three choices of parameter c.

It is intuitively plausible that items of all three kinds may exist in practice, depending on the difficulty and other properties of the item. (In particular the second case may correspond to an item whose solution depends on a sudden insight, the occurrence of which is independent of how long the person has been at the item so far.) Thus, the distribution which results from substituting experssion (2.4) in equations (2.2) and (2.3) will be quite a flexible one which can model a wide variety of types of items depending on the value of the parameter c, which is hence a crucial one.

Making the stated substitutions and carrying out the integration called for, we obtain

$$(2.5) \qquad F(t) = \begin{cases} 1 - \exp\left[-\left(\frac{t - t_o}{\mu_o}\right)^c\right] & \text{for } t \geq t_o \\ 0 & \text{for } t < t_o \end{cases}$$

as the Weibull distribution function and

$$(2.6) \qquad f(t) = \begin{cases} \frac{c}{\mu_o}\left(\frac{t - t_o}{\mu_o}\right)^{c-1} \exp\left[-\left(\frac{t - t_o}{\mu_o}\right)^c\right. & \text{for } t \geq t_o \\ 0 & \text{for } t < t_o \end{cases}$$

as the Weibull density function. In the system-reliability theory literature the three parameters t_o, μ_o, and c are referred to as the location, scale, and shape parameters, respectively. Since we let t_o be the lower limit of the range of t in the general derivation of F(t) from h(t) above, it is clear that this parameter is the theoretical value of t such that prob$(t<t_o) = 0$. Thus it is natural to call this the location parameter. The scale parameter μ_o specifies the $100(1-e^{-1})$ percent point of the distribution of $t - t_o$ [i.e., prob $(t<t_o+\mu_o) = 1 - e^{-1} \approx .632$] as may readily be verified by letting $t = t_o + \mu_o$ in the expression for the distribution function F(t) given in equation (2.5). The shape parameter c is the most interesting of the three, for it determines the general shape assumed by the density function. If $c \leq 1$, there is no mode and the density function decreases monotonically with t. If $c > 1$, the distribution is unimodal and skewed, with mode at $t_o + \mu_o(1-1/c)^{1/c}$. Interestingly, the skewness changes from positive to negative at approximately c = 3.60. Figure 2 shows the density functions of Weibull distributions with $t_o = 2$, $\mu_o = 15$, and four selected values of c.

The mean and variance of a random variable following the Weibull distribution $W(t_o,\mu_o c)$ are as follows:

$$(2.7) \qquad E(t) = t_o + \mu_o \Gamma(1+1/c)$$

and

$$(2.8) \qquad Var(t) = \mu_o^2[\Gamma(1+2/c) - \Gamma^2(1+1/c)],$$

where $\Gamma(\cdot)$ is the gamma function, defined as

$$f(t) = (c/\mu_0{}^c)\,(t-t_0)^{c-1} * \exp\left[-\left((t-t_0)/\mu_0\right)^c\right]$$

c	μ_0	t_0
.7	15	2
1	15	2
1.5	15	2
2.5	15	2

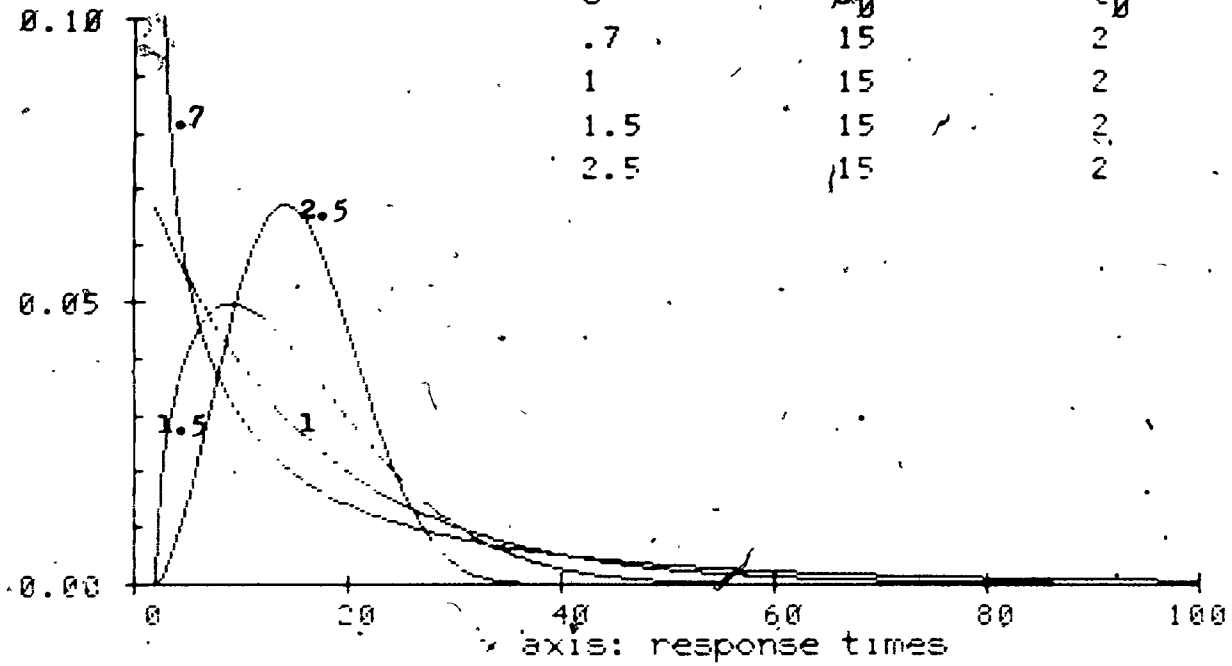


Figure 2. Weibull density functions with t_0=2, μ_0=15, and four values of c

$$\Gamma(m) = \int_0^\infty e^{-u} u^{m-1} du$$

$$= (m-1)! \qquad \text{when } m \text{ is an integer.}$$

2.2 Comparison of Several Related Distributions

As a matter of incidental interest as well as a possible aid in subsequent discussions relating the value of the shape parameter c to the nature of the item or other unit of a test, we display the density functions of several related (or in some sense similar) distributions and also indicate what each of them reduces to when c = 1.

The density function of the two-parameter gamma distribution used by Rasch (1960) as a model for the distribution of time taken to read a passage of N words is, in a notation consistent with what we are using for the Weibull distribution,

$$(2.9) \qquad f_{2g}(t) = [\frac{1}{\mu_o}(\frac{t}{\mu_o})^{c-1} \exp\ (-\frac{t}{\mu_o})]\ /\Gamma(c).$$

The N in Rasch's equation (6.6) corresponds to our c, and his λ to our $1/\mu_o$. Equation (2.9) is equivalent also to that given by Restle and Davis (1962) in their k-stage model for problem solving, where k = c. To serve in either the Rasch or the Restle-Davis model, the c in equation (2.9) must thus be an integer, but there is no such requirement in the density function (2.6) of the Weibull distribution. If we let c = 1 in equation (2.9), we get the density function for the one-parameter negative exponential distribution:

$$(2.10) \qquad f_{1e}(t) = \frac{1}{\mu_o} \exp\ (-\frac{t}{\mu_o}).$$

Again, $1/\mu_o$ is customarily written as λ and called the intensity parameter.

On the other hand, if we let c = 1 in equation (2.6), we get the density function of the two-parameter negative exponential distribution

$$(2.11) \qquad f_{2e}(t) = \frac{1}{\mu_o} \exp\left(-\frac{t - t_o}{\mu_o}\right),$$

which is the model found by Brée (1975) to offer a better fit to the distributions of solving times for Restle and Davis' three problems than did the two-parameter gamma distribution, (2.9). Brée called (2.11) the negative exponential distribution with shift in location. In other words, this density function starts at $t = t_o$ instead of $t = 0$ as does the one-parameter negative exponential distribution, (2.10).

Now it is well known that when c is an integer greater than 1, the two-parameter gamma distribution (2.9) is a c-fold convolution of the one-parameter negative exponential distribution (2.10). In other words, if there are c independent random variables $t_1, t_2, \ldots, t_c$ each following the one-parameter negative exponential distribution (2.10) then their sum $t = t_1 + t_2 + \ldots + t_c$ follows the two-parameter gamma distribution (2.9). [Thus Rasch's model for the distribution of reading times for an N-word passage amounts to saying that the reading times for each word follow a negative exponential distribution and that the N distributions are statistically independent. Similar remarks hold for Restle and Davis' k-stage problem-solving model.]

In analogy to the fact just stated, that the two-parameter gamma distribution (for integer c) is a c-fold convolution of the one-parameter negative exponential distribution, it might be tempting to jump to the conclusion that the Weibull distribution (2.6) with integer c (>1) is a c-fold convolution of the two-parameter negative exponential distribution (2.11)--in view of the fact that (2.6) reduces to (2.11) when c = 1. This is not the case, however. Rather, a c-fold convolution of the two-parameter negative exponential distribution gives rise to the three-parameter gamma distribution having the density function

$$(2.12) \qquad f_{3g}(t) = \left[\frac{1}{\mu_o}\left(\frac{t - t_o}{\mu_o}\right)^{c-1} \exp\left(-\frac{t - t_o}{\mu_o}\right)\right]/\Gamma(c).$$

Note that letting c = 1 in this equation also leads to equation (2.11). Thus, the Weibull distribution and the three-parameter gamma distribution have in common the property that they both reduce to the two-parameter negative exponential distribution when c = 1.

A comparison of equations (2.6) and (2.12) shows that the density functions of the Weibull and the three-parameter gamma distributions are strikingly similar. Aside from the absence of the normalizing constant $1/\Gamma(c)$, (2.6) differs from (2.12) only in the presence of

c at two places where it is lacking in (2.12). Thus, in a sense, the Weibull distribution may be regarded as a somewhat generalized form of the three-parameter gamma distribution.

3. PARAMETER ESTIMATION

It must be conceded that the methods we used to estimate the parameters of the Wiebull distribution and those of the two-parameter gamma were not the best possible. But, operating as we were under tight time constraints and since PLATO IV uses a time-sharing mode with rather limited storage capacity allocated to any one user, we had to make do with relatively simple methods with reasonable accuracy. It might be mentioned in passing that PLATO V terminals, each equipped with a microcomputer of its own, are becoming more and more widely available, and they would circumvent much of the limitations under which we operated.

It was unfortunate also that we did not become cognizant of the three-parameter gamma distribution early enough to include it among the distributions to be fitted to our data. However, a comparison of equations (2.6) and (2.12) suggests that not much has been lost, since the two density functions are remarkably similar, as noted earlier, and if anything the Weibull distribution appears to have a slight edge on the gamma in flexibility. Whether this is indeed so must await future research, however.

3.1 The Weibull-Distribution Parameters

The problem of estimating the parameters of a Weibull distribution has been the subject of a number of papers (e.g., Harter and Moore, 1965; Johns and Lieberman, 1966; Mann, 1967, 1969; Lemon, 1974). Most of these, however, either deal with two-parameter versions of the Weibull distribution (i.e., when one or another of the three parameters is assumed known) or present iterative methods whose programming appears to be an enormous job. Believing that the maximum likelihood method would be the most accurate, and before becoming familiar the papers just cited (since Sato [1971] spoke only of a rudimentary method using a special Weibull probability paper) one of the present authors derived the likelihood equations and struggled for some time to solve them. He concluded--correctly as it turned out--that they were capable of solution only by tedious, iterative methods. Since time was of the essence, he abandoned the maximum likelihood approach[2] and improvized a rough-and-ready method based on linear regression, as follows.

---

[2]He subsequently became aware that a computer program listing for this method could be obtained from H. L. Harter, Aerospace Research Laboratories, Wright-Patterson AFB, Ohio. But even adopting and implementing this on the PLATO system would have been a lengthy task.

We first rewrite equation (2.5) for the Weibull distribution function as

$$1 - P = \exp\left[-\left(\frac{t - t_o}{\mu_o}\right)^c\right],$$

where F(t) has been denoted by P for short, it being understood that it is a function of t and corresponds to the observed proportion of examinees who respond to an item by a given time t. Taking the natural logarithms of both sides of this equation gives

$$\ln(1-P) = -\left(\frac{t - t_o}{\mu_o}\right)^c.$$

Changing the signs of both sides and taking their natural logarithms again yields

$$(3.1) \qquad \ln\ln(1-P)^{-1} = c\,\ln(t-t_o) + \ln(\mu_o^{-c}).$$

If we now let

$$(3.2) \qquad \ln\ln(1-P)^{-1} \equiv Y,$$

$$(3.3) \qquad \ln(t-t_o) \equiv X$$

and

$$(3.4) \qquad \ln(\mu_o^{-c}) \equiv a,$$

equation (3.1) becomes

$$(3.5) \qquad Y = cX + a,$$

which looks just like an ordinary linear regression equation of Y on X. The only (but big) difference is that X itself is not completely observable, because it depends on one of the unknown parameters t_o, as equation (3.3) shows. [Note that if we were dealing with a two-parameter Weibull distribution with t_o known (usually, $t_o = 0$) then (3.5) would indeed be a regular linear regression equation, and the estimation of c and a would be a simple matter.]

Therefore, we had to resort to a trial-and-error method to estimate t_o first, and then apply the standard methods of linear regression to estimate c and a, from which in turn μ_o is determined via equation (3.4). The principle adopted for guiding the trial-and-error procedure was to maximize the correlation between $Y = \ell n \ell n (1-P)^{-1}$ (which is observable) and $X = \ell n(t-t_o)$, which becomes an observable once some value is given to t_o. The search started by dividing the interval $[0, t_{min} + t_{min}/200]$ into 20 subintervals (where t_{min} is the smallest observed response time) and calculating r_{xy} with t_o given trial values equal to the endpoints of these subintervals. Next, the (closed) interval between the trial value of t_o yielding the largest value for r_{xy} and the adjacent one giving the next largest value for r_{xy} was divided into ten equal subintervals and their endpoints were taken as the second set of trial values for t_o with which to calculate r_{xy}. Finally, the interval between the trial values among this set that yielded the two largest values for r_{xy} was again divided into ten subintervals and their endpoints were taken as the third set of trial values t_o with which to calculate r_{xy}. The optimal among these trial values was taken as our final estimate $\hat{t}_o$ of t_o.

Once the estimate $\hat{t}_o$ is determined, $X = \ell n(t-\hat{t}_o)$ is calculable for each observed value of t, and thence $\hat{c}$ is computed from

$$\hat{c} = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sum X^2 - (\sum X)^2/n},$$

where n is the number of observed response times. Then $\hat{a}$ is computed as $\hat{a} = \overline{Y} - \hat{c}\overline{X}$, and $\hat{\mu}_o$ is obtained by solving equation (3.4) for it:

$$\hat{\mu}_o = (\exp \hat{a})^{-1/\hat{c}}.$$

This completes our estimation of the three Weibull parameters, rough-and-ready though it is. In the subsequent sections we omit the circumflexes and write t_o, μ_o and c for these estimates to simplify the notation, since we will not need to refer to the true parameter values.

3.2 The Gamma-Distribution Parameters

Here, too, the maximum likelihood estimates would probably have been the most desirable, but due to the limitations already mentioned we again adopted a simpler method, the method of moments in this case. Since only two parameters have to be estimated, it suffices to express the theoretical mean and variance in terms of the parameters and equate these expressions to the observed mean and variance, respectively.

The required expressions, computable from equation (2,9), are

$$E(t) = \mu_o c$$

and

$$Var(t) = \mu_o^2 c.$$

It readily follows that

$$\mu_o = Var(t)/E(t)$$

and

$$c = [E(t)]^2/Var(t).$$

Hence,

$$\hat{\mu}_o = s_t^2/\bar{t}$$

and

$$\hat{c} = (\bar{t})^2/s_t^2$$

may be taken as estimates for μ_o and c, where $\bar{t}$ and s_t^2 are the sample mean and variance, respectively. In the sequel, we write α for $\hat{c}$ and β for $\hat{\mu}_o$ to avoid confusion with the corresponding Weibull parameter estimates.

4. DESCRIPTION OF DATA

As mentioned earlier many different data sets were used in this study. Some of them were from lessons (which include instructional

segments, exercises, and quizzes), tests (pre and post) on matrix algebra that were developed by one of the present authors for the dual purpose of serving as a self-study course to accompany several statistics courses taught by the other author and two of his colleagues in the Educational Psychology and Psychology Departments of the University of Illinois at Urbana-Champaign and for gathering data for this study. Others came from over 30 lessons, and their accompanying tests, on general and special vehicle maintenance training developed by the Chanute Air Force Base Computer-Based Education (CBE) Project Group under the sponsorship of the Advanced Research Projects Agency (ARPA) of the Department of Defense.

4.1 Matrix Algebra Lessons and Tests

The matrix algebra course, written on the PLATO system by one of the present authors with some assistance from one of her associates, is intended for graduate students in educational and psychological statistics--particularly multivariate statistics--who do not have much mathematical background. Topics covered include the basic definitions and simple operations of matrix algebra, matrix multiplication, matrix inversion (including the definition and calculation of determinants), linear transformations and axis rotations, and eigenvalue problems. The course is divided into five lessons corresponding to the above topics, and their average completion times range from 20 minutes to 2 hours per lesson. (See Appendix A for several sample pages of the course.)

The PLATO system permits a student to make any number of passes through any instructional unit, which may be the actual instructional segment, a set of exercises, or a quiz, and which is called an "area" in PLATO terminology. Each area is identified by the lesson number preceded by the letter i, e or q (for instruction, exercise or quiz), and followed by the instructional segment number within that lesson. Thus, for example, "i036" refers to the sixth instructional segment in lesson 3, while "e036" refers to the exercise set for the sixth instructional segment in lesson 3. An exception occurs in lesson 5, which contains but one instructional segment as such (i051) followed by three exercise sets (e051, e052 and e053) of the problem-solving type to augment the instruction. There are 36 areas in all, whose codes and content matter are listed in Table 1.

The set of data for any area includes, among other things, the name or ID number of each student who went through that area completely at least once, the pass (or try) number, and the time he/she took on each pass. For the purposes of data analysis, only the time taken on the first pass (if completed) through each area for each student was considered.

Table 1

List of Areas and their Content in Matrix Algebra Lessons

Introduction to Matrices

area	content
i011	Definitions and simple operations of a matrix
i012	Use PLATO as a calculator
e011	Eleven exercises

Matrix Multiplications

i021	Multiplication of two matrices A and B
e021	Foru exercises
i022	Multiplication is not commutable ,i.e., $AB \neq BA$
e022	Four exercises
i023	Scalar product
e023	Four exercises
i024	Matrix product
e024	Four exercises
i025	Quadratic product
e025	Four exercises
i026	The principles of matrix operations
e026	Exercises
i027	Diagonal matrices
e026	Four exercises
i028	Scalar matrices and identity matrix
e028	Four exercises

Determinant and Inversion of a Matrix

i031	Identity matrix
q031	Five item quiz
i032	Definition of the determiant of a matrix
i033	Evaluation of the determinant of a matrix
q033	Five item quiz
i034	Cofactors, expansion of a determinant
e034	Exercises for cofactors, expansion of a determinant
i035	Properties of determinants
i036	Adjoint and inverse of a matrix A

Matrix and Linear Transformations

i041	An example of linear transformation;axis rotation
i042	Properties of orthogonal transformations
i043	SSCP matrix

Eigenvalues and eigenvectors

i051	Definition of eigenvalues and eigenvectors
e051	Calculate eigenvalues
e052	Calculate eigenvectors
e053	Normalization of eigenvectors

A 48-item pretest was given to all students before they studied the matrix algebra course. (See Appendix B for a list of the items, plus a sample item as it appears on the PLATO screen.) The original version of this pretest was constructed a year and a half before Fall 1976, and had been used in the multivariate statistics course. It was designed to minimize guessing by permitting students who did not know the subject matter related to a given item to omit it and go on to the next by pressing the NEXT key without having to choose any of the multiple-choice options in the earlier item. There were 88 students who tried every item and were thus likely to have taken the test seriously in an earnest desire to find out their initial level of knowledge. The data for these 88 students are referred to as the "pre-revision data" in the sequel. After all 48 items have been answered, the pretest provides feedback by indicating which option the student chose for each item, the correct option for that item and, at the very end, a recommendation as to which lesson the student should start from.

In Fall 1976, a revision of the pretest was undertaken in light of information obtained from the original version. Some displays, wordings and options were changed, but the biggest change was that the NEXT key could no longer be used without choosing some option in each item, thus forcing students to respond to every item. The feedback system was retained in the revised version, however. Data from the new version of the pretest are referred to as the "post-revision data" below.

At the same time, a posttest for the first two lessons combined (simple operations and matrix multiplication) and one for each of the other lessons (lesson 3, matrix inversion; lesson 4, transformations; and lesson 5, eigenvalues and eigenvectors) were implemented and the time and performance scores on these posttests have been collected since then. Only those who completed each lesson could take the corresponding posttest.

Since most instructors of the relevant statistics courses did not forcibly require all students in their classes to study the matrix algebra lessons on PLATO, data for these lessons came mainly from volunteers who selected the topics according to their own judgment. But taking the pretest was requested by most instructors. Thus, computer-managed instruction (CMI) was not carried out, and instead of forcing the students to adopt a predetermined strategy, almost complete freedom of choice of learning strategy was allowed the students. We therefore did not develop a computer-managed router of the mastery learning type. Instead, data collection routines were implemented within the lessons and tests so that all the students' behavioral records were collected. That is, for each student and each area, the time spent in that area, the number of questions attempted (whether in an instructional segment, an exercise or a quiz) the number of questions ultimately answered correctly, the number of questions correctly answered on the first try, and the number of times the student requested and received on-line

help, were recorded. In addition, for the items in the quizzes, the pretest and the posttest, more detailed data were collected: the response time for each item, whether the item was answered correctly or not, and the number of times the item was attempted.

4.2 Chanute AFB CBE Project Lessons and Tests

These lessons have been developed over a period of more than 10 years, as a cooperative enterprise between the Chanute AFB CBE Project group and members of the Military Training Center (MTC) group at the Computer-Based Education Research Laboratory (CERL) of the University of Illinois at Urbana-Champaign, for the purpose of training special and general purpose vehicle repairpersons (Dallman, 1977). There are 34 lessons, comprising about 30 hours of instruction, along with a criterion-referenced test for each. The lessons are homogeneous in subject matter (in the sense that they do not naturally form a hierarchically organized set) and tutorial in style for the most part. Nevertheless, they are arranged in a specific order and students must achieve mastery in one lesson as assessed by the end-of-lesson test before they can proceed to the next. If mastery is not achieved, they must repeat the lesson. A listing of the contents of the lessons is given in Appendix C.

The 34 associated tests consist mostly of matching and multiple-choice items, and they vary from 5 to 20 items in length. Only one pass is allowed through each test and no feedback is given. The tests are called MVE (for Master Validation Exams) and are numbered to correspond to the lessons; e.g., the test given at the end of lesson 101 is denoted MVE 101. The mastery levels are set at 80 percent, but the cutoffs actually used are somewhere between 75 and 90 percent correct.

A lesson is said to be validated when 90 percent of the students have achieved mastery by getting 75 to 90 percent of its MVE test items correct. The samples yielding the data for analysis in this study consisted of about 30 students per lesson, though not necessarily the same 30 each time. No modifications of lessons were made until all the students finished them, and all lessons were validated (after which they might be modified) between April and September, 1975, inclusive.

The data collected included test scores on the MVE tests, completion time for each test, the completion time for each lesson each time it was studied (which may be just once or several times, depending on how quickly mastery was achieved), and the total time spent on each lesson until mastery. The last mentioned time is called the "mastery time" for each lesson in the sequel. Unlike the matrix algebra lessons, data are available only for entire lessons and not for their constituent parts.

A flow chart of the lessons and tests in the Chanute AFB CBE Project is shown in Figure 3.

4.3 Teaching Strategies and Lesson Styles

Since the matrix algebra course and the Chanute AFB CBE Project course in motor vehicle maintenance differ considerably in their teaching strategies and lesson styles, we compare them here although some of the descriptions were already given above.

Virtually every lesson in the Chanute course followed the simple tutorial learning activity that can be characterized as a linear series of instructions and questions. Every student is required to proceed through the same material in each lesson regardless of prior knowledge or ability. Since these students were first-year Air Force draftees with only a high-school education for the most part, this lesson style is probably well suited for them. They probably could not be trusted with much freedom of choice.

By contrast, each lesson in the matrix algebra course has an index page at the beginning, as illustrated in Figure 4. Each student can choose a particular lesson segment covering the topic of his/her choice. Since the students taking the matrix algebra course were all graduate students in educational psychology, psychology or accountancy (with a few from other departments) they were bright and motivated enough to control their own learning activities, and hence this lesson style was probably the best for them.

It should be noted that, in both courses, the posttest scores were significantly higher on the average than were the pretest scores, thus permitting us to infer that learning did take place regardless of which lesson style was used.

To make somewhat more detailed comparisons, in the matrix algebra course some topics are taught by drill and practice strategies while others are taught by problem-solving strategies. The particular strategy chosen was adapted to the nature of the topic. For instance, simple subject matter such as matrix addition and multiplication are taught with the aid of exercises, following the instructional segments, that are designed to give students practice in calculations, while more difficult material such as eigenvalues and eigenvectors are taught with the help of exercises of the problem-solving type. All but one lesson contained the provision of allowing the student to go back for review to the preceding frame within any area (instructional segment, exercise or quiz), and also to go clear back to the index page (see Figure 4). In the latter case the student could choose to go to an area other than that in which he/she was working before re-calling the index. This resulted in some messy data which had to be discarded in our analysis,

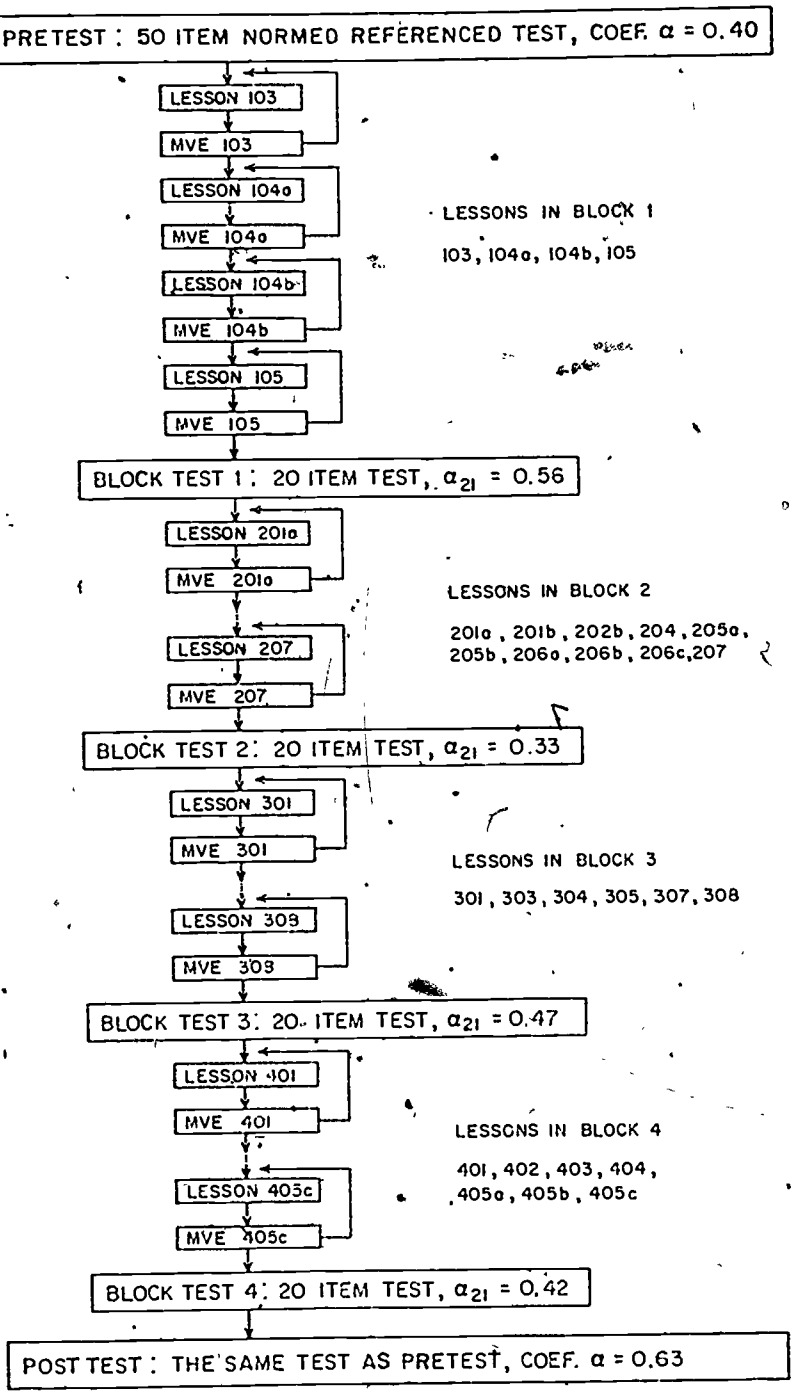


Figure 3

Block diagram of student flow in Chanute AFB CBE Project

MULTIPLICATION OF MATRICES

You can enter the section you worked last by typing the section number. If this is your first time in this lesson you should begin from 2.1

Fig. 4 Index page of the lesson teaching multiplication of matrices

for the times spent in the two areas became fuzzy. (In fact, it led to a reduction from about n = 300 to about n = 100 in some analyses.) However, since many students requested this option, it was implemented after the Fall 1976 semester. Such are the disadvantages of collecting data in conjunction with learning activities in which the freedom to which graduate students are accustomed is permitted!

By contrast, if a student did not achieve mastery at the end of a Chanute lesson, he was required to repeat the entire lesson. The total time taken by each student to master each lesson was recorded for the purpose of lesson validation. It was this mastery time that was used for our analyses of the Chanute data. The situation here was much "cleaner" and under strict control by the instructor in typical military style.

Finally, it should be mentioned that some students in both courses took notes during their studying, which of course lengthened their study times. Since the percentage of such students amounted to only about 1 or 2 percent of the total sample, we did not discard the data from these students--which would have been difficult to do without closer monitoring and log keeping than we were able to effect. We rationalized this state of affairs by regarding note taking as part of the normal learning strategy for some people, and hence the time for this activity should be included in their study time.

5. ANALYSIS OF PREREVISION DATA

Before presenting the results of analyses based on the prerevision data (which, it will be recalled, are the data from the matrix algebra pretest prior to its revision in Fall 1976) we give a brief description of the PLATO IV system and the programs that were implemented on it for this study. Virtually all of the programs were written by Robert Baillie aside from some contributions made by Tamar Weaver, Kay Tatsuoka, and Jerry Dyer in this order of involvement.

5.1 The PLATO IV System and the Programs

PLATO IV (Programmed Logic for Automated Teaching Operations) is a computer-based education system developed at the University of Illinois at Urbana-Champaign having a large-scale central computer (the Control Data Corporation Cyber 73-74) with about 1,000 terminals connected by telephone lines throughout the United States. Approximately 5,000 hours of instructional material have been used in several hundred subject-matter areas, and additional lessons are constantly being developed. The target populations range from preschool children to

graduate students, including such diverse groups as prison inmates and special adult-education recipients like industrial workers, Armed Forces personnel, and the physically handicapped.

The PLATO system itself was used as the primary analytic tool for analyzing the student data collected automatically on the system, besides serving as the deliverer of instructional material. Data processing can be done directly without having first to punch the data onto cards, and the results can be utilized for such diverse purposes as adaptive testing, computer-managed instruction and item analyses leading to modification of weak instructional units.

The computer language used is called the TUTOR, which is somewhere between FORTRAN and assembler language in its capability and precision for numerical work. Each word of the computer is 60 bits in lingth, which provides for greater accuracy than most existing computers. This feature is especially at a premium when iterative calculations are required as in the computing of gamma or beta integrals which abound in statistical work. Approximation routines for various theoretical distributions were written, along with that for the Kolmogorov-Smirnov test of goodness-of-fit of observed with theoretical distributions. This involved a great deal of adaptations and modifications of existing statistical programs, mainly from the IBM Scientific Subroutine Package (IBM, 1972).

Since PLATO operates on a large-scale time-sharing mode, special problems exist in programming for it that are similar to using a minicomputer in terms of storage size. The core size per user is limited to 1650 words at a time for data processing. If the computational requirement exceeds this limit, transfer routines must be developed for moving the data and intermediate results back and forth between the disk storage and the core, where data processing is done successively within the limit. A list of the computer programs written on PLATO expressly for this study is given in Table 2.

5.2 Weibull Fitting of Item Response-Time Data

The 48 items of the matrix algebra pretest shown in Appendix B were implemented in the test frame written by James Kraatz of CERL and modified by K. Tatsuoka into two parts--one allows us to edit data and the other stores and transforms the data format so as to be acceptable by the programs for estimating the Weibull and gamma parameters. Data editing was necessary for several reasons. One was that we were interested only in the first pass data, as mentioned earlier, even though second-pass and third-pass data were also on record. Sometimes the system would "crash" while the student was taking the test and he/she had completed, say, the tenth item. Then the response-time record for the remaining 38 items would consist of blanks. The TUTOR would

Table 2

List of PLATO Programs Developed under the Project

Program Name	Brief Description
matx4	The 48 item matrix algebra test. It collects performance and response time data.
edittest	Shows and allows us to edit the data. Do the simple item analysis.
storetest	Transforms and stores the data collected from matx4 into a permanent storage (dataset).
datam	Calculates the item characteristics of 48 items and estimates the individual student's performance level.
gram	Estimates the individual gain scores by regressing the time score difference onto the pre-test, post-test, and other variables.
Kappa	Calculates Kappa index from a test.
subr	Calculates various probability functions.
matsubr	Calculates the determinant of a matrix, inverse, eigen values, and eigen vectors.
cutoff	Evaluates the optimum cutoff scores of a criterion-referenced test, estimates the probabilities of false positive and negative.
llab	Plots various relationships between the test information such as $2I$ vs. probability of false positive, etc.
wb2	Estimates Weibull parameters of the data from mat24.
statedit	Input output routine with a data format that was adapted as the standard format for all programs developed by the NIE project.
wb2area	Estimates Weibull parameters from the data stored via statedit format.
kolmo	Kolmogorov-Smirnov testing routine for matx4-data format.
gamma	Kolmogorov-Smirnov testing for statedit format

(Table 2 cont.)

wgraf	Comparison of Weibull distributions associated with the items. Density functions of various Weibull parameters. Plotting of conditional response rates.
kgraf	Draws graphs of Weibull distribution and density function based on typed-in parameters.

Note. Various univariate and multivariate statistics routines were developed to analyze our on-line data stored on the PLATO system. Also several transformation programs were developed. Their descriptions and main programmers are listed in Appendix C.

memorize the item number at which the crash occurred, and would automatically send the student to the eleventh item on his/her second entry. At that time the response-time data for the first 10 items would be blanks and actual times would be recorded from the eleventh item on. We then had to combine the two sets of data to get the score and response-time data for the first try for that student. Sometimes we would encounter data records in which the same response option was chosen for all items, thus indicating that the student (or instructor) was merely examining the items and not taking the test. Such data would, of course, have to be deleted. All told, there was about a 20 percent attrition due to editing to clean up the data.

Using these cleaned-up data, a Weibull distribution was fitted to the observed time-score distribution of each item in three ways: once for the entire sample, secondly for the subgroup of students who answered the item correctly (called the "OK subgroup") and finally for the students who got the item wrong (called the "NO subgroup"). The fit of the observed to the theoretical distribution was tested each time by the Kolmogorov-Smirnov test of goodness of fit. The OK subgroup and NO subgroup had considerably different estimated Weibull parameters, but both showed very good fits for most items. Ninety-three and 92 percent of the 48 items had p-values for the Kolmogorov-Smirnov test of goodness of fit with Weibull distributions that exceeded .20 in the OK and NO subgroups, respectively, and 65 and 83 percent exceeded .50 respectively. Considering the fact that two items which needed corrections during the fall semester of 1976 due to unclear display on the screen or ambiguous wording showed very poor fit, with p-values of 0.0053 and 0.0550, the fit of nearly all of the other items is seen to be satisfactory to excellent. Weibull distributions did not fit the time-score data of the total sample as well as they did those of the two subgroups. Only 69 percent of the items had p-values larger than .20, and 56 percent had values larger than .50. This fact suggests that students in the two subgroups are going through different processes to complete each item; thus the nature of the time-score data in the two groups might be entirely different.

Tables 3 and 4 show the p-values and the maximum discrepancies (z) for the OK and NO subgroups, respectively. Tables 5 and 6 show the estimated Weibull parameters for the 48 items in the two subgroups.

Next are shown figures illustrating the degree of observed to theoretical distribution fits for two typical items in each of the two subgroups (Figures 5 through 8). The fits (or lack thereof) are shown in two ways: first by superimposing the observed cumulative distribution graph onto the theoretical curve with the estimated parameters; and second by fitting the regressions lines of $\ell n \ell n (1-P)^{-1}$ on $\ell n (t-t_o)$ to the observed scatterplot after determining the value of t_o yielding the maximum correlation between these two quantities (see Section 3.1).

Table 3

Kolmogorov-Smirnov Tests for Matrix Algebra Pretest; OK Group

item	p	z	N	item	p	z	N
1)	Ø.2ØØ7	1.ØØ617	77	25)	Ø.3489	Ø.9329	62
2)	Ø.1571	1.1277	82	26)	Ø.3956	Ø.8979	59
3)	Ø.8832	Ø.5853	68	27)	Ø.6966	Ø.7ØØ8	48
4)	Ø.9716	Ø.4871	69	28)	Ø.582Ø	Ø.777Ø	44
5)	Ø.25Ø4	1.ØØ188	79	29)	Ø.4534	Ø.8579	33
6)	Ø.4419	Ø.8656	81	3Ø)	Ø.6424	Ø.741Ø	34
7)	Ø.4675	Ø.8486	67	31)	Ø.8145	Ø.635Ø	25
8)	Ø.1444	1.1463	61	32)	Ø.9983	Ø.3873	18
9)	Ø.416Ø	Ø.8834	67	33)	Ø.5176	Ø.8164	28
1Ø)	Ø.6237	Ø.752Ø	69	34)	Ø.8719	Ø.5942	18
11)	Ø.3829	Ø.9Ø72	22	35)	Ø.5414	Ø.8Ø17	3Ø
12)	Ø.9621	Ø.5Ø28	27	36)	Ø.5891	Ø.7727	38
13)	Ø.6196	Ø.7545	42	37)	Ø.3821	Ø.9Ø78	27
14)	Ø.9918	Ø.4335	46	38)	Ø.8945	Ø.5759	24
15)	Ø.42Ø5	Ø.88Ø3	54	39)	Ø.7887	Ø.6522	3Ø
16)	Ø.6378	Ø.7437	63	4Ø)	Ø.9963	Ø.4Ø81	18
17)	Ø.8898	Ø.5798	28	41)	Ø.864Ø	Ø.6ØØ2	17
18)	Ø.2979	Ø.9749	59	42)	Ø.9714	Ø.4875	31
19)	Ø.8424	Ø.616Ø	29	43)	Ø.9726	Ø.4852	24
2Ø)	Ø.3747	Ø.9133	4Ø	44)	Ø.8142	Ø.6355	21
21)	Ø.574Ø	Ø.7818	37	45)	Ø.6191	Ø.7548	22
22)	Ø.ØØ184	1.5314	59	46)	Ø.9776	Ø.4754	12
23)	Ø.7264	Ø.69Ø8	28	47)	Ø.9954	Ø.4143	15
24)	Ø.21Ø1	1.ØØ611	47	48)	Ø.9884	Ø.4467	7

Pretest for all subjects before 1976 Fall semester; 'goodness of fit' testing for Weibull distributions

Table 4

Kolmogorov-Smirnov Tests for Matrix Algebra Pretest;NO Group

item	p	z	N	item	p	z	N
1)	Ø.9944	Ø.4288	9	25)	Ø.2918	Ø.9818	15
2)	Ø.9996	Ø.3536	2	26)	Ø.7675	Ø.6656	19
3)	Ø.7178	Ø.6965	17	27)	Ø.8488	Ø.6177	27
4)	Ø.9941	Ø.4222	16	28)	Ø.2981	Ø.9817	33
5)	Ø.8236	Ø.6291	6	29)	Ø.5341	Ø.8862	44
6)	1.0000	Ø.2887	3	30)	Ø.9817	Ø.5697	43
7)	Ø.4598	Ø.8537	16	31)	Ø.6827	Ø.7178	52
8)	Ø.9673	Ø.4946	22	32)	Ø.9174	Ø.5553	54
9)	Ø.7671	Ø.6659	15	33)	Ø.9884	Ø.4689	46
10)	Ø.8998	Ø.5728	14	34)	Ø.7752	Ø.6687	57
11)	Ø.Ø594	1.3268	61	35)	Ø.6533	Ø.7345	39
12)	Ø.8388	Ø.6191	56	36)	Ø.7389	Ø.6881	36
13)	Ø.9958	Ø.4116	41	37)	Ø.5578	Ø.7916	47
14)	Ø.9998	Ø.3438	37	38)	Ø.5712	Ø.7835	45
15)	Ø.7183	Ø.7886	27	39)	Ø.6857	Ø.7153	43
16)	Ø.9525	Ø.5164	17	40)	Ø.8435	Ø.6153	54
17)	Ø.5822	Ø.8262	54	41)	Ø.1867	1.2185	53
18)	Ø.7438	Ø.6887	21	42)	Ø.5588	Ø.7959	48
19)	Ø.5998	Ø.7668	52	43)	Ø.9164	Ø.5562	42
20)	Ø.9359	Ø.5363	34	44)	Ø.3885	Ø.9658	47
21)	Ø.9493	Ø.5285	35	45)	Ø.1186	1.2838	45
22)	Ø.8638	Ø.6818	28	46)	Ø.9442	Ø.5268	58
23)	Ø.Ø827	1.2628	58	47)	Ø.8813	Ø.5868	39
24)	Ø.6237	Ø.7521	31	48)	Ø.7991	Ø.6454	22

Pretest for all subjects before 1976 Fall semester; 'goodness of fit' testing for Weibull distributions

Table 5

The Three Weibull Parameters for Matrix Algebra Test Items

Items	t_0	m.c.	c	μ_0
1.	2.71	Ø.98	1.Ø5	39.52
2.	1.59	Ø.98	1.32	21.98
3.	1.78	Ø.99	1.15	16.86
4.	1.46	Ø.99	1.34	21.91
5.	2.51	Ø.99	1.26	12.84
6.	2.77	Ø.99	1.15	11.ØØ
7.	2.63	Ø.97	1.14	31.46
8.	6.Ø4	Ø.99	1.27	36.62
9.	1.63	Ø.99	1.38	28.37
1Ø.	3.52	1.ØØ	1.38	18.28
11.	1Ø.53	Ø.97	Ø.91	29.82
12.	Ø.ØØ	Ø.99	1.34	1Ø9.63
13.	8.18	Ø.99	1.25	64.46
14.	9.14	Ø.99	1.44	38.17
15.	15.52	Ø.99	Ø.98	61.57
16.	5.47	Ø.99	1.16	38.47
17.	Ø.ØØ	Ø.98	1.17	139.65
18.	2.33	Ø.99	1.52	16.89
19.	1.67	Ø.98	1.17	48.25
2Ø.	Ø.ØØ	Ø.96	1.45	66.83
21.	Ø.ØØ	Ø.98	1.13	68.66
22.	3.19	Ø.98	1.59	19.82
23.	2.71	Ø.97	1.Ø1	16.86
24.	1.74	Ø.98	1.19	12.93
25.	2.75	Ø.98	1.Ø9	8.ØØ
26.	1.65	Ø.99	1.21	7.69

Table 5 (con t)

The Three Weibull Parameters for Matrix Algebra Test Items

Items	γ_0	m.c.	c	μ_0
27.	1.96	0.99	0.75	19.08
28.	1.95	0.99	0.77	16.39
29.	3.32	0.99	0.83	26.15
30.	2.74	0.97	0.96	24.36
31.	6.60	0.99	0.96	24.34
32.	2.79	0.99	0.70	35.26
33.	0.00	0.95	1.24	175.50
34.	7.68	0.98	0.77	33.73
35.	0.67	0.99	1.92	59.48
36.	2.95	0.99	0.78	12.49
37.	4.73	0.99	1.02	14.34
38.	2.66	1.00	1.06	13.32
39.	2.68	0.99	1.28	8.42
40.	3.75	0.99	0.70	8.09
41.	15.85	0.98	1.14	25.48
42.	0.46	1.00	1.47	25.81
43.	1.29	1.00	0.98	17.20
44.	2.80	1.00	0.89	12.86
45.	2.65	0.99	1.08	10.94
46.	2.08	0.99	0.94	22.52
47.	3.68	1.00	0.88	15.56
48.	6.56	0.99	0.81	10.72

Pretest given before 76 Fall semester, OK subgroup

Table 6

The Three Weibull Parameters for Matrix Algebra Test Items

items	t_0	m.c..	c	μ_0
1.	10.66	0.99	0.88	33.96
2.	0.00	1.00	0.98	3.58
3.	0.00	0.97	1.06	13.05
4.	1.73	0.99	0.69	14.54
5.	0.00	0.99	1.07	6.02
6.	0.90	1.00	0.55	2.13
7.	1.52	0.97	0.78	38.31
8.	0.00	0.99	1.05	37.07
9.	0.73	0.98	0.85	15.16
10.	0.00	0.98	0.92	20.69
11.	1.35	0.96	1.11	31.80
12.	0.59	1.00	0.91	69.67
13.	4.49	1.00	1.01	38.08
14.	0.00	1.00	1.13	46.24
15.	1.52	0.98	0.66	60.92
16.	0.92	0.99	0.38	19.33
17.	0.72	0.99	0.88	51.96
18.	1.86	0.99	0.80	15.80
19.	0.26	0.98	0.92	25.56
20.	0.92	0.99	0.62	22.99
21.	0.92	0.99	0.69	34.93
22.	0.18	0.99	1.04	13.43
23.	2.69	0.99	1.23	14.41
24.	0.00	0.97	1.56	12.48
25.	0.56	0.99	1.01	13.18
26.	0.64	0.99	0.76	17.01

Table 6 (con't)

The Three Weibull Parameters for Matrix Algebra Test Items

items	t_0	m.c.	c	μ_0
27.	Ø.73	Ø.99	Ø.95	9.32
28.	Ø.94	Ø.99	Ø.68	6.21
29.	Ø.63	Ø.99	Ø.87	13.67
30.	Ø.85	Ø.99	Ø.93	18.Ø7
31.	Ø.66	1.ØØ	Ø.95	22.63
32.	1.Ø3	Ø.99	Ø.94	51.93
33.	1.77	Ø.99	Ø.8Ø	58.73
34.	Ø.55	Ø.99	1.15	31.48
35.	Ø.94	Ø.99	Ø.64	18.6Ø
36.	Ø.71	Ø.99	1.ØØ	7.22
37.	Ø.52	Ø.99	Ø.98	16.3Ø
38.	Ø.95	Ø.99	Ø.85	1Ø.57
39.	Ø.87	1.ØØ	Ø.92	7.77
4Ø.	Ø.78	1.ØØ	1.1Ø	8.34
41.	Ø.32	Ø.99	1.16	39.Ø3
42.	1.97	Ø.99	Ø.75	11.66
43.	Ø.44	Ø.99	Ø.97	9.67
44.	Ø.88	Ø.99	Ø.75	5.99
45.	Ø.96	Ø.99	Ø.74	6.34
46.	Ø.7Ø	1.ØØ	Ø.8Ø	11.7Ø
47.	Ø.ØØ	Ø.99	Ø.82	17.1Ø
48.	Ø.ØØ	Ø.99	Ø.96	24.23

Pretest given before Fall 76 semester, NO subgroup

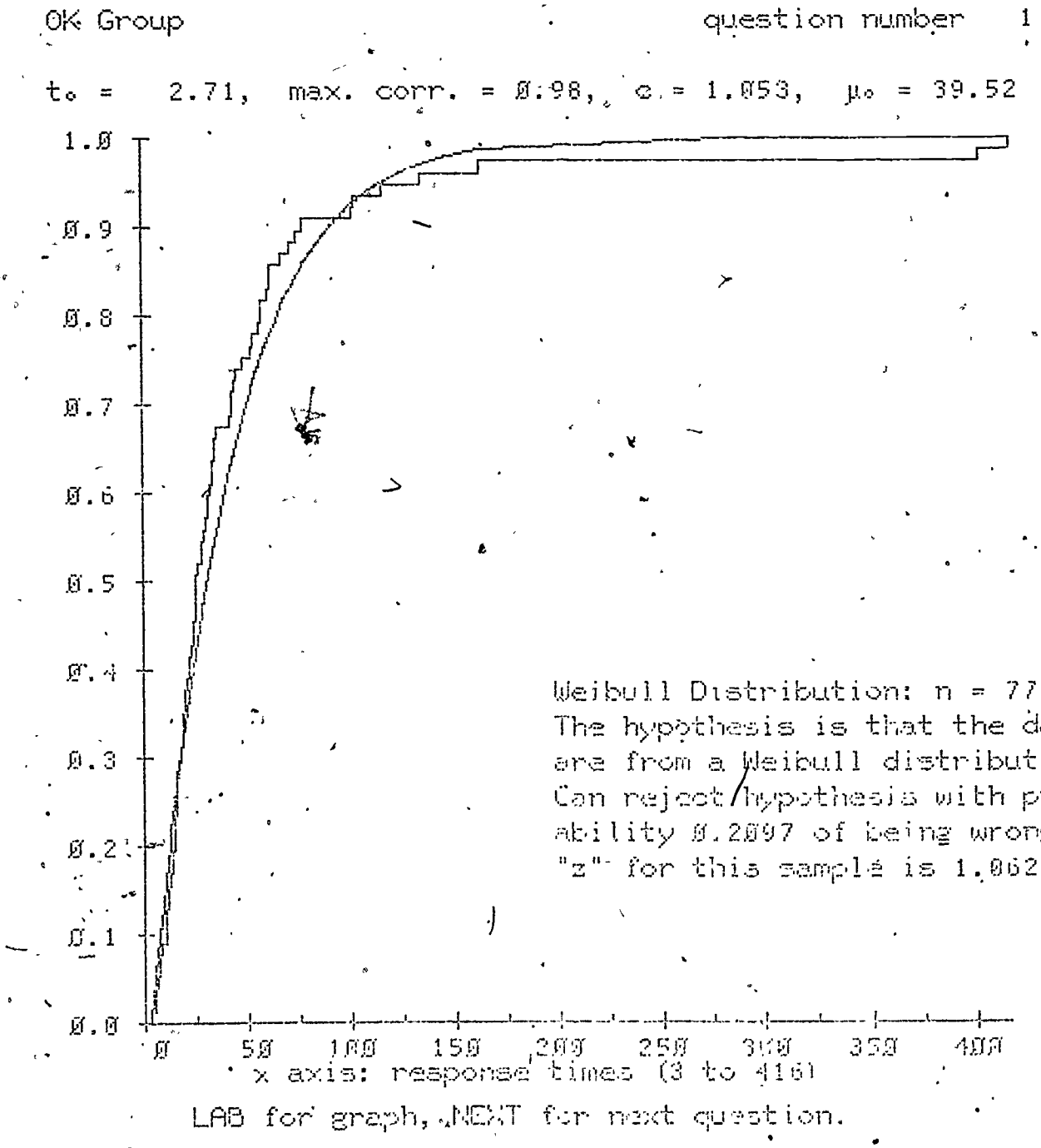
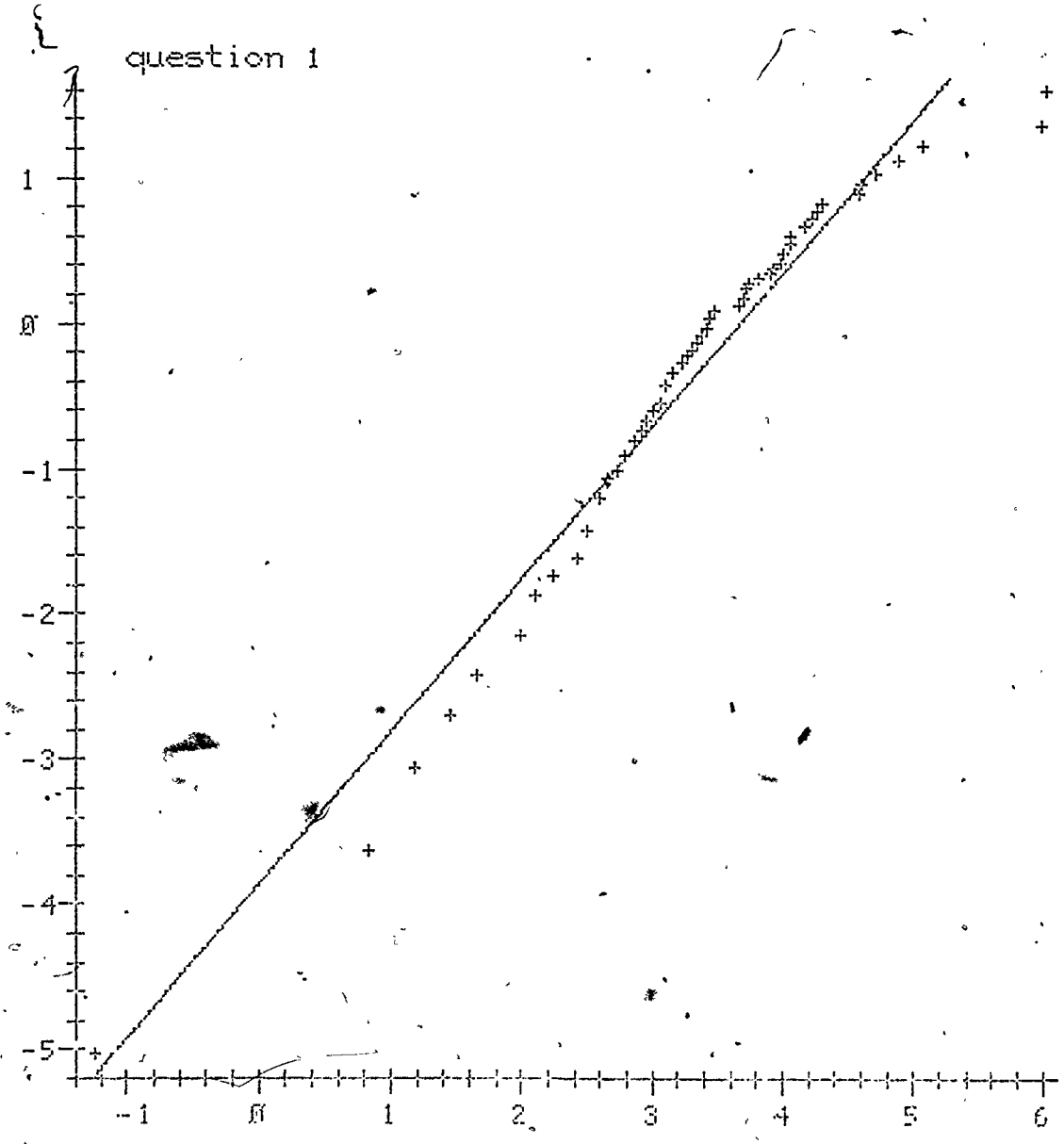


Figure 5 Goodness of fit test for the time-score data and Weibull distribution function



n=77, x(1)=3, x(n)=416, max. correl.=0.98, p=0.2097
t_0= 2.71, c=1.053, y_0=39.52, $\Sigma(dev^2)/(n-1)$=0.0036

NEXT: next question, LAB: reset slope and intercept

Figure 6 Plot of $\ln\ln(1-P)^{-1}$ as yaxis and $\ln(t-t_0)$ as x axis. The line is the regression equation of y on x.

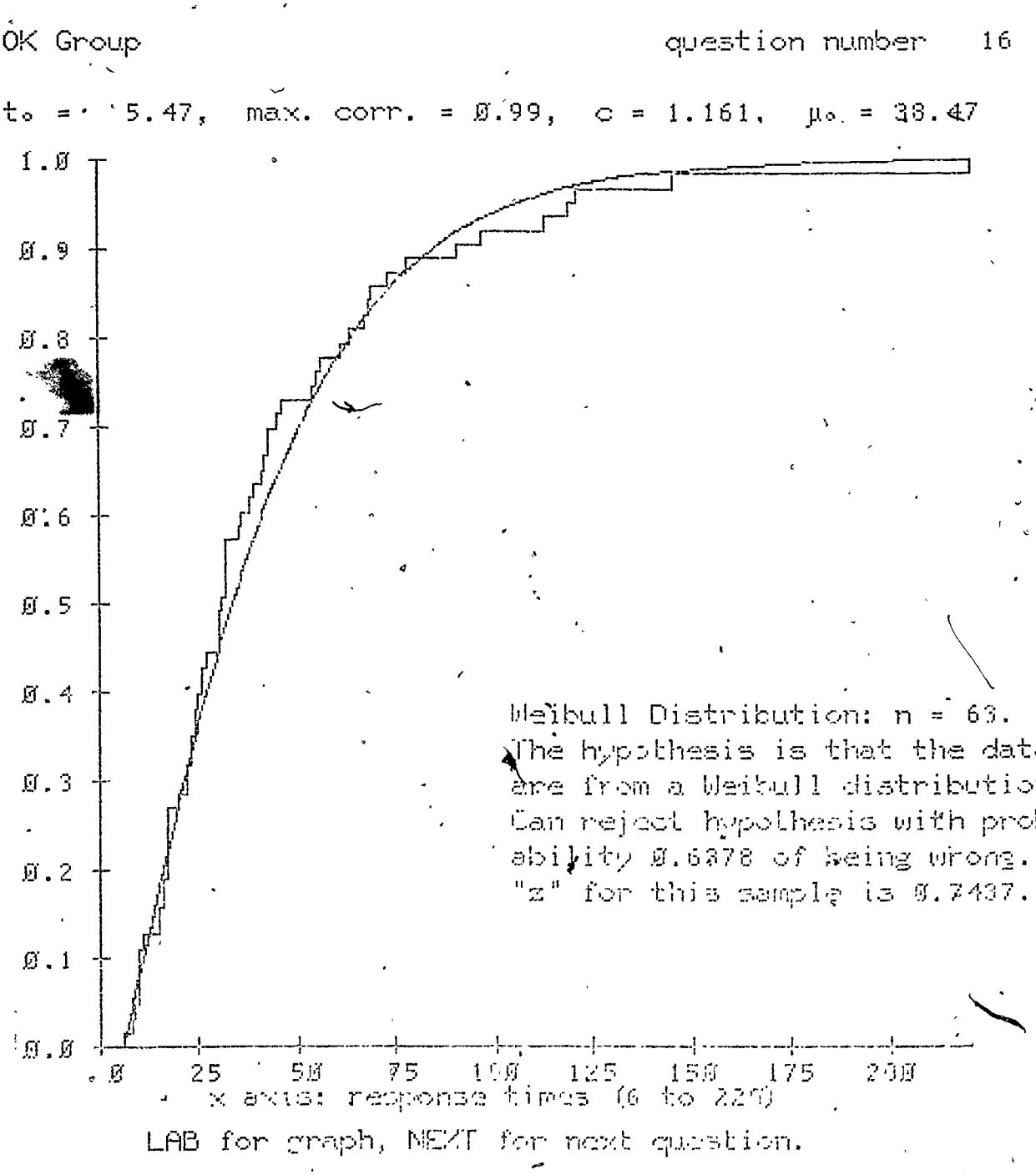
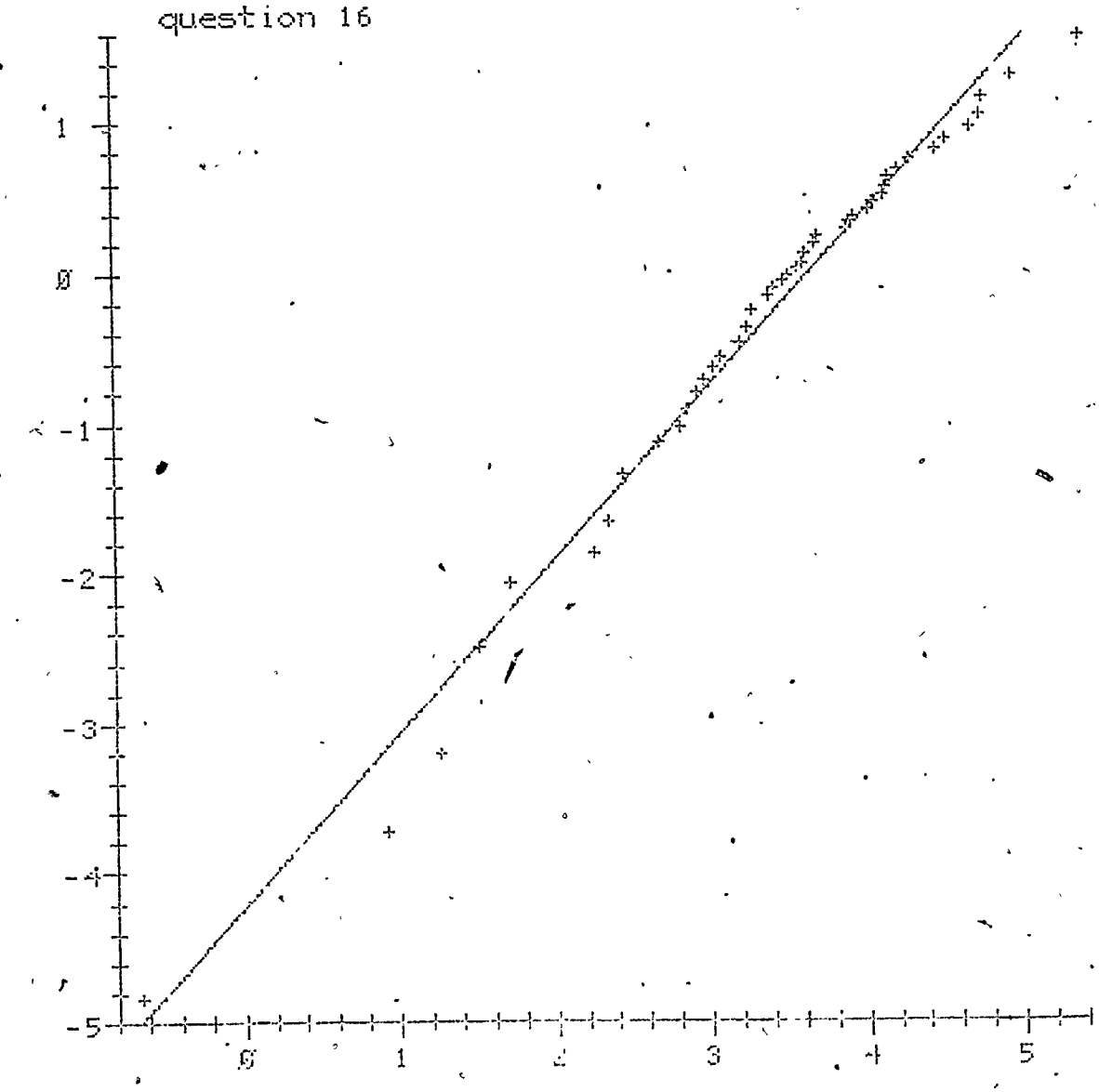


Figure 7 Goodness of fit test for the time-score data and Weibull distribution function



n=63, x(1)=6, x(n)=220, max. correl.=0.99, p=0.6378
t_o= 5.47, c=1.101, μ_o=38.47, $\Sigma(dev^2)/(n-1)$=0.000015

NEXT: next question, LAB: reset slope and intercept

Figure 8 Plot of ln $\ln(1-P)^{-1}$ as y axis and $\ln(t-t_0)$ as x axis. The line is the regression equation of y on x.

Since the Kolmogorov-Smirnov testing procedure looks only at the maximum discrepancy (z) between the observed and theoretical cumulative distributions, the resultant "goodness-of-fit" depends partly on the size of the intervals or units of measure used. In the matrix algebra pretest, the item response times were recorded to the nearest second, and hence this undesirable feature of the Kolmogorov-Smirnov test seldom manifests itself there. Even so, when the time range is small (as in true-false items) and the sample size is relatively large there are occasions when the p-value is rather small despite the fact that the fit looks very good to the eye. This trouble (if indeed it be a trouble) increases when we come to analyze lessons, where the times taken are recorded only to the nearest 10 seconds and is further aggravated when we get to the Chanute data, where the time unit is minutes and some lessons take only 25 minutes at a maximum.

5.3 Characteristics of the Pretest Items

The performance-score data from the pretest items were processed by the computer program developed for computing various item parameters. (See description in Appendix D.) Some of the results are displayed in Table 7. The first column in this table, labeled "Difficulty 1," shows the values of the traditional difficulty index--i.e., proportion of subjects getting the items right. The second column ("Difficulty 2"), on the other hand, gives values of a modified difficulty indes due to Loeschner (personal communication). It is defined as the estimated average probability that the particular item (i) is answered correctly but another item (j) is answered incorrectly when a randomly drawn subject is given both items i and j. The formula is

$$(5.1) \qquad \text{Difficulty 2} = \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{n_{ij}}{(n-1)N},$$

where n is the number of items in the test,
N is the number of subjects taking the test, and
n_{ij} is the number of subjects who got item i right but item j wrong.

The reason we regard this alternative "difficulty" (actually "facility") index worth considering along with the traditional difficulty index is as follows. The topics covered in a matrix test are, by their nature, hierarchically ordered (or, more strictly speaking, linearly related). For instance, in order to be able to compute a matrix inverse, one must know what an identity matrix is, must know how to multiply matrices, must know what cofactors and adjoints are, and how to calculate the determinant of a (square) matrix. These prerequisite knowledges must have been mastered earlier and the required calculations

Table 7

Difficulty indecis, discriminating power, and the correlation of the total score and item-time score

item	Difficulty1	Difficulty2	$r_{s,i}$	$r_{s,ti}$
1	.90	.44	.43	-.03
2	.95	.48	.48	.03
3	.79	.39	.32	.08
4	.80	.39	.40	.07
5	.92	.46	.39	-.07
6	.94	.47	.45	.10
7	.78	.36	.50	.02
8	.71	.32	.55	.16
9	.78	.36	.53	.08
10	.80	.37	.59	.05
11	.26	.13	.07	.06
12	.31	.13	.35	.39
13	.49	.20	.49	.37
14	.54	.23	.50	.28
15	.63	.27	.62	.14
16	.73	.33	.58	.13
17	.33	.11	.58	.39
18	.69	.30	.64	.19
19	.38	.12	.53	.27
20	.48	.18	.60	.24
21	.44	.17	.56	.41
22	.69	.31	.50	.30
23	.33	.13	.43	.35
24	.55	.23	.56	.33
25	.72	.32	.61	.20
26	.69	.32	.45	.01
27	.56	.23	.58	.38
28	.51	.22	.48	.21
29	.38	.14	.59	.38
30	.40	.16	.50	.38
31	.29	.11	.44	.39
32	.22	.08	.36	.38
33	.33	.11	.60	.47
34	.21	.08	.39	.34
35	.35	.14	.41	.46
36	.44	.18	.51	.32
37	.31	.12	.43	.42
38	.28	.11	.41	.35
39	.34	.12	.57	.42
40	.21	.07	.41	.41
41	.20	.07	.40	.42
42	.36	.13	.62	.46
43	.28	.10	.52	.43
44	.24	.09	.41	.38
45	.26	.07	.65	.32
46	.14	.03	.55	.34
47	.17	.08	.19	.42
48	.22	.07	.54	.07

carried out without error in order to achieve the goal of getting a matrix inverse. These prerequisites were taught in the first three lessons of the matrix algebra course. The various test items in the pretest were roughly ordered by the difficulty of the topic involved.

Bob Linn (personal communication) suggested that the difficulty index (proportion of correct answers) of items should be expected to have a perfect negative rank-order correlation with difficulty of topic. This is so because, for instance, nobody should be able fully to understand the import of the identity matrix without first knowing how to multiply matrices. Thus, if item i tests for the knowledge of matrix multiplication, while item j tests for understanding the concept of the identity matrix, it is natural to expect that anyone who got item j correct will also have answered item i correctly.

Now, we sorted our subjects x items score data by item "difficulty" (proportion of subjects answering correct) and by total score earned by each subject. The result is a plot of dots and blanks in a pattern like that shown in Figure 9, which resembles a scalogram. The upper left-hand corner represents the score (dot = 1, blank = 0) on the easiest item earned by the highest-scoring subject. Then the points at which the number of dots to the left equals the total score were connected by a "step line," which Sato (1977) called the "S-curve." If the data were perfect, i.e., if the items were scalable in Guttman's (1947) sense, and the item scores were error-free, then the sum of the estimated conditional probabilities $p(X_j = 1 \mid X_i = 1)$ over $j = i + 1, i + 2, \ldots, 48$ (where X_i and X_j are the scores on items i and j, which are either 1 or 0) will be given by the shaded area in the figure. This value is associated with the relative importance of item i to the items testing for more advanced topics. We related these sums for the 48 items with the difficulty of the topic being tested and found a nearly perfect rank-order matching; only three items were disarrayed. (Thanks are due to Bob Linn for suggesting that we consider the conditional probabilities. It was our idea to sum them.)

It should now be clear that Difficulty 2 is related to the sum of the conditional probabilities complementary to those considered above. Hence this alternative "difficulty" index is of interest quite apart from the traditional difficulty index.

The discriminating power $r_{s,i}$ in the classical test theory sense is shown in the third column, while the correlation $r_{s,ti}$ between item response time and total test score is given in the last column. Since the test is a pretest for a difficult subject that requires higher cognitive skills and most of the students were not mathematics or physical-science majors, almost 65 percent of the items were tough problems. The 88 students on the basis of whom the results in Table 7 were

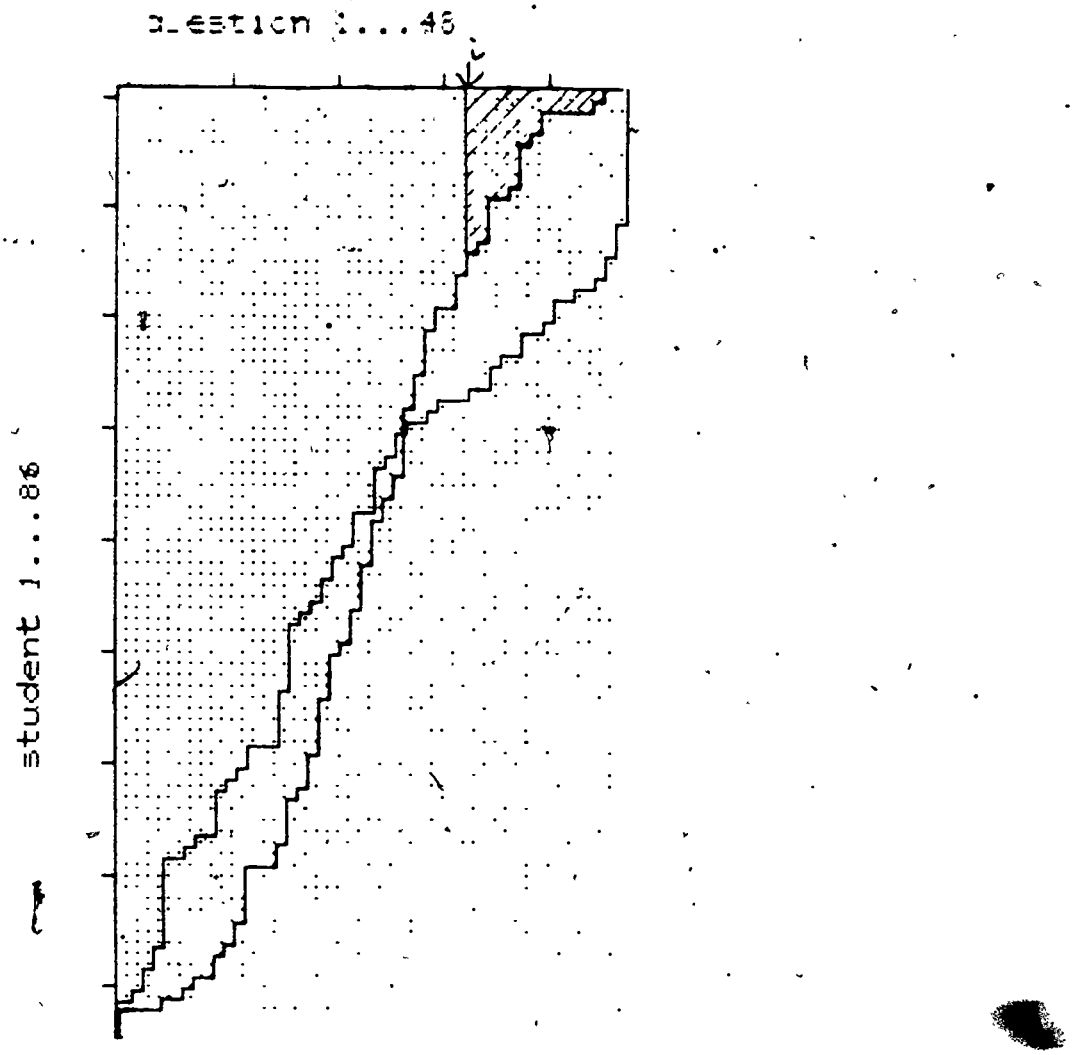


Figure 9 Graphical explanation of difficulty2 (Loeschner's facility index)

Note. The 48 items are ordered by their difficulty index so that the easiest item is placed at the leftmost, and the hardest at the rightmost position. The 86 students are also ordered by their scores from the top line as the highest, to the bottom as the lowest. A small dot stands for a correct answer 1, and a blank is for wrong, 0. The two step lines represent the total scores of each student and number of right answers for each item respectively. The shaded area represents difficulty2 of item i when a test is perfect.

computed excluded those who were taking (or who had completed) the most advanced of the three statistics courses using the matrix algebra course, and hence the $r_{s,i}$ values are not very high. It may be noted that $r_{s,ti}$ increases as the difficulty indices get smaller, which is reasonable because a person who gets higher test scores tends to stick longer with difficult items while less able students give up on them and go on to the next item.

Table 8 shows the correlations among the four measures (two of which are themselves correlation coefficients) that were displayed in Table 7 and were discussed above. It is seen that the two types of item difficulty (actually "facility") indices correlate almost perfectly with each other. (It might therefore be argued that the second index, Difficulty 2 is gratuitous, but it does have some desirable properties discussed above that are not possessed by the traditional difficulty index.) On the other hand the two "discrimination indices" are uncorrelated with each other, and instead $r_{s,ti}$ (or rather its Z transform) shows a moderate negative correlation with the "difficulty" indices.

Table 8

Correlations Among Two Item Difficulty Indices and Two Discriminating Power Measures

	1.	2.	3.	4.
1. Difficulty 1	1			
2. Difficulty 2	.991	1		
3. $Z(r_{s,i})^*$	.241	.175	1	
4. $Z(r_{s,ti})^*$	-.492	-.544	-.018	1

*These are Fisher's Z-transforms of the correlation coefficients shown in the parentheses.

Recalling from Table 7 that no item actually had negative $r_{s,ti}$ values to speak of (only two values were negative, but they were practically zero), a low $r_{s,ti}$ value means that students having high total scores and those with low total scores showed little difference in time taken

to respond to that item. Hence, the moderate negative correlations between $r_{s,ti}$ and the "difficulty" indices, just noted, imply the following relations: It was the easy items (i.e., those with large "difficulty" index values) that tended, by and large, to exhibit little differences in response time between those with high and low prior knowledge of matrix algebra. Conversely, the more difficult items tended to show larger differences in response time between high and low total score students, with high scoring students tending to take longer time. We may therefore infer that students with higher prior knowledge of matrix algebra tended to persevere longer on difficult items while those with low prior knowledge tended to give up on them sooner. This is a reasonable result, and by itself is almost trite (except that it does seem to confer some construct validity to the test) but it has some implications for subsequent interpretations of the Weibull shape parameter, c.

The Relation between Discriminating Power and Time. Woodbury (1963) and Novic (1966) developed a model involving time that identifies the measurement process with the realization of a stochastic process. However, their definition of the time parameter, t, is the examiner-controlled time allowed for the test, in other words, the length of the test, whereas the time score we have been using is the time taken by an examinee as needed. Their studies showed that there is some optimum time that maximizes the reliability of a test. Even though their definition of time for a test and its relationship with test theory are quite different from our usage, we were convinced that by controlling time after the fact in the scoring process, we could demonstrate a similar relation from our data between time score and some established concept in test theory. Testing out our hunch, we found two interesting empirical relations between time and discriminating power. Specifically, when a test item is easy, there is an optimal time point within a relatively short time interval such that the discriminating power of the item becomes the largest. On the other hand, for difficult items, the longer the time allowed the better the discriminating power. These relations were observed fairly consistently for 48 items in two samples of about 80 and 100 subjects--i.e., data from the prerevision and postrevision matrix algebra pretests--and also for the posttests for the lessons on multiplication, matrix inversion, transformations, and eigenvalues and eigenvectors.

Figures 10 through 12 illustrate these relations, while Table 9 displays the numerical detail on which Figure 10 is based. (Corresponding tables for Figures 11 and 12 are omitted to save space.) To

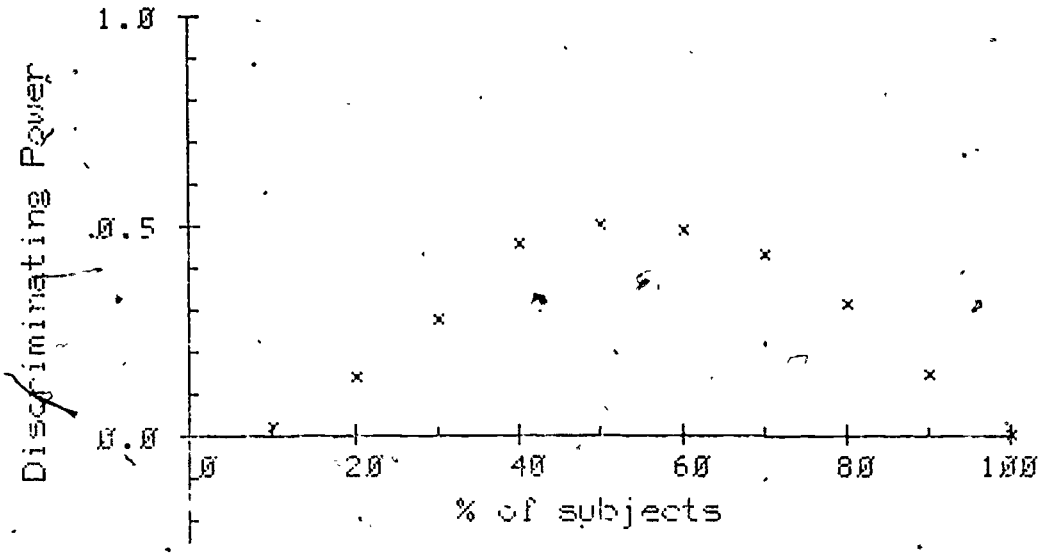


Figure 10 Discriminating powers over 10 cutoff times (in seconds) for OK subgroup

Table 9

10 Points in Figure 12, item 20

%*	cut time	N	r**
1Ø	2Ø	4	Ø.Ø18
2Ø	29	8	Ø.139
3Ø	33	12	Ø.272
4Ø	37	16	Ø.456
5Ø	42	2Ø	Ø.497
6Ø	54	24	Ø.487
7Ø	67	28	Ø.432
8Ø	82	32	Ø.311
9Ø	11Ø	36	Ø.145
1ØØ	1Ø8	4Ø	Ø.ØØØ

* % of subjects, ** discriminating power

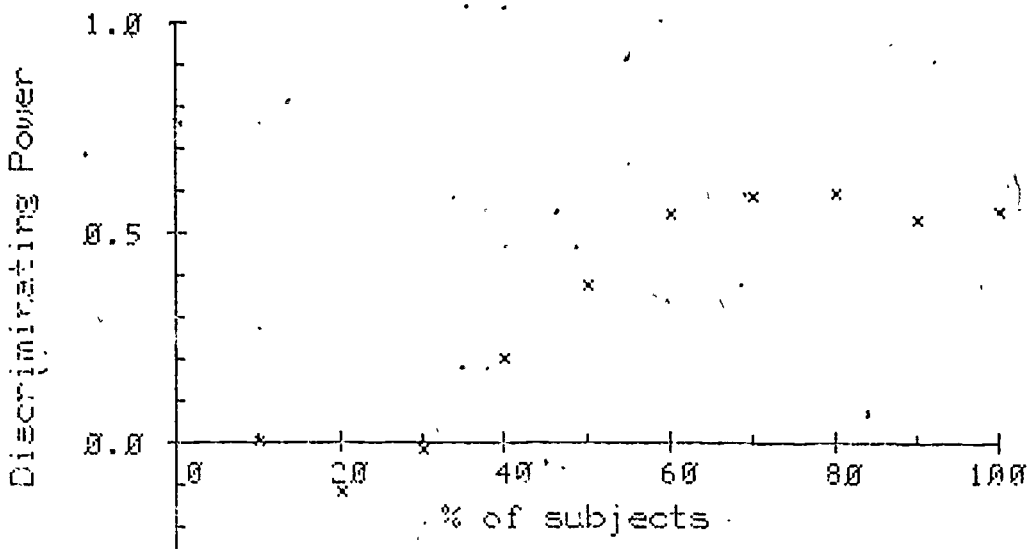


Figure 11 Discriminating powers over 10 cutoff times (in seconds) for all subjects The item is question 20 which is the same item in Figure 10.

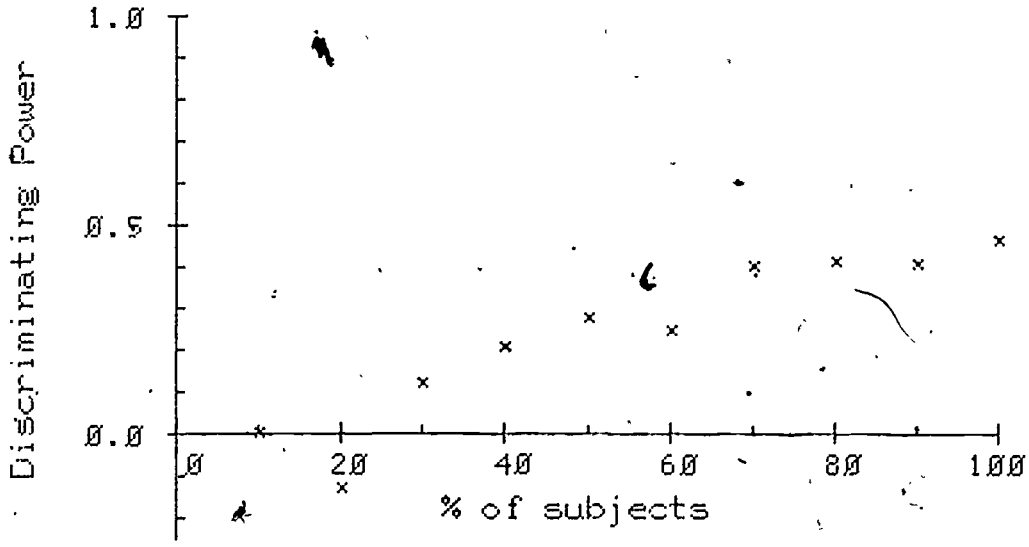


Figure 12 Discriminating powers over 10 cutoff times (in seconds) for all subjects. The item is question 16.

explain the construction of Figure 10 and the contents of Table 9, the 40 subjects in the OK subgroup for Item 20 were first arranged in ascending order of their times taken. The fastest 10 percent (n = 4) of the subjects, with response times no greater than 20 seconds, were taken and only these subjects were regarded as having answered Item 20 correctly. The Item-20 score and the total test score for 36 subjects were thus modified, with the scores (1 or 0) on the other items left unchanged for all 40 subjects. The point biserial correlation calculated between the modified Item-20 score and the modified total score for the 40 subjects, is what is shown as the first entry, .018, in the last column ("adjusted discriminating power") of Table 9 and is the ordinate of the first point plotted in Figure 10. Next the fastest 20 percent, with response times no greater than 29 seconds, were scored 1, and the others scored 0, on Item 20 and the total scores accordingly modified. The point biserial thus calculated is the second entry, .139, of the last column in Table 9 and is ordinate of the second point in Figure 10. The same process is repeated for the remaining cutoff percentages, 30, 40, . . ., 90 percent, yielding adjusted discriminating powers .272, .456, . . ., .145, respectively. The last cutoff percentage (100 percent) necessarily yields a point biserial value of zero, because all 40 subjects now are scored 1 on Item 20, since only the OK subgroup was used. For this subgroup Item 20 was obviously an extremely easy item (everyone got it right), and the maximum adjusted discriminating power .497 occurs when the cutoff percentage is 50 percent, with a cutoff time 42 seconds, thus illustrating the time vs. discriminating power relation stated above for an easy item.

Figure 11 shows the time vs. discriminating power relation for the same item, but now using the total sample of 74 "earnest" subjects (i.e., those who chose an option at all for Item 20). Of course not all of the fastest 10 percent (n = 7) were scored 1 on Item 20 this time but only those among the seven who actually got the item right were so scored. Similar scoring was used for the fastest 20 percent, fastest 30 percent, etc. through the entire group. Item 20 is now a moderately difficult item, with 40 out of 74 subjects getting it right, and the maximum discriminating power, .597, now occurs at cutoff percentage 80 percent with cutoff time 70 seconds.

Figure 12 presents an exception to the rule. Item 16 was an easy item (73 percent got it right, as shown in Table 7), and yet the maximum discriminating power, .462, occurs with 100 percent cutoff. Thus, the empirical generalization stated earlier is not a perfect one, suggesting that other factors besides item difficulty must affect the relation between discriminating power and time. Theoretical work on this issue is planned.

5.4 Interpretation of Weibull Parameters

We saw in Section 5.2 that the response-time data for practically all of the 48 matrix-algebra pretest items were well fitted, and those for a large majority of them were excellently fitted, by Weibull distributions. It is now time to engage in some interpretations of the observed fit. The first thing to note is that the Weibull distribution for an item in the OK subgroup (those who got the item correct) and that in the NO subgroup showed interesting differences. This is apparent from a comparison of Tables 5 and 6, given earlier. Let us now focus on a couple of specific items and compare their Weibull parameters for the two subgroups.

For example, Item 10, which asks for the transpose of a 2 x 2 matrix (see Appendix B), shows quite a contrast between the two sets of Weibull parameters. The OK subgroup has larger values for all three basic parameters than does the NO subgroup:

	t_o	c	μ_o	μ
OK subgroup	3.52	1.33	29.82	30.73
NO subgroup	0.00	.92	20.69	21.50

Here μ is the theoretical mean, denoted earlier by E(t) and related to the three basic parameters through equation (2.7):

$$\mu = t_o + \mu_o \Gamma(1+1/c).$$

Similarly, Item 16 (finding the product of a (2x3) and a (3x2) matrix) has Weibull parameters as follows:

	t_o	c	μ_o	μ
OK subgroup	5.47	1.16	38.47	39.94
NO subgroup	.92	.38	19.33	29.85

Since t_o is the theoretical minimum time required for examinees to arrive at their answer, it is only natural that the NO subgroup had the smaller value for both items. Most members of this subgroup simply pressed the NEXT key or made an incorrect guess. They usually don't know what the transpose of a matrix is or how to multiply matrices. They may have had some exposure to the rudiments of matrix algebra in a college algebra course a long time ago, but since they had no

further contact with matrices they have forgotten what little they knew. Thus it seems safe to infer that anyone whose response time to an item is closer to the t_o value for the NO subgroup than to the t_o for the OK subgroup must have guessed at the answer instead of attempting to solve the problem. However, this is still in the realm of speculation, and we will examine the issue further in the context of posttest data.

Thirty-seven out of the 48 items have larger values of c (the shape parameter) in the OK subgroup than in the NO subgroup, but the opposite is true for 11 items. Six of these 11 items were of the true-false type, and two involved either an ambiguity of wording or an inconspicuous symbol (' for the transpose of a matrix). Thus, a majority of the items for which the c value in the NO subgroups was larger than in the OK subgroup had something unusual about them.

Returning to the two items cited above, both were among the 37 "normal" items for which the OK subgroup had the larger value of c. Looking at the μ values for the two items, we can infer that Item 10 was easier than Item 16. (In fact Table 7 shows that the "difficulty index"--which should be called the "ease index"--had the values .802 and .733 for the two items, respectively.) The difficulty index and c are indeed positively correlated, at least for the OK subgroup.

Table 10 in the next subsection shows that the c for the OK subgroup correlates (across the 48 items) .41 with the difficulty index[3] as computed for the total sample, while the c for the NO subgroup correlates -.15. These are not exactly high correlations, but when the c is based on the total sample (not shown in Table 10) the correlation increases to .56. If the c from the NO subgroup is partialled out, the partial correlation between c in the total sample and the difficulty index is .70. Since the time-score distribution in the total sample does not fit the Weibull distribution as well as those in the OK and NO subgroups separately, however, the parameter c based on the total sample may not be very meaningful. We may have to introduce a composite Weibull distribution (cf. Mann, Schafer and Sigpurwalla, 1974, pp. 140-142) to fit the total sample, but we have not done so in the present study.

The upshot of the foregoing discussions is that the shape parameter c has something to do with item difficulty, but not so much as to be identified with it. In a sense, c has a "richer" meaning than the

---

[3]Unless the suffix '2' is attached, "difficulty index" will always denote the traditional difficulty index, and not the alternative index introduced by Loeschner.

usual concept of difficulty, since it determines the shape of the cumulative distribution curve. The nature of its relationship with the distribution shape is illustrated in Figure 13, which depicts the Weibull distribution with $t_o = 10$, $\mu_o = 30$, $c = 1.5$ (curve 1) and that with $t_o = 10$, $\mu_o = 30$, $c = .8$ (curve 2). Curve 1 is seen to approach its asymptote more rapidly than curve 2 does.[4] Although, by definition, the graph of any distribution function must asymptote to $F(\cdot) = 1$, it may approach different values within "reasonable" ranges of the time variable, thus indirectly reflecting different item difficulty levels.

To further illustrate how c determines distribution shape with real item data, we again return to Item 16. Figure 14 shows the distribution curves for Item 16 in both the OK subgroup (curve 1) and the NO subgroup (curve 2). Curve 1 starts at $t_o = 5.47$ on the time axis and converges to 1 faster than does Curve 2. It is interesting to note that about 40 percent of the NO-subgroup examinees leave this item before the (theoretical) minimum time, 5.47 seconds, for the OK subgroup. Ten percent of the NO-subgroup examinees spent too long a time without achieving success while almost all subjects in the OK subgroup arrived at the answer in 130 seconds. These facts suggest that it is not necessary to allow more than 130 seconds for people to answer Item 16. The density-function curves for both subgroups are also shown in Figure 14, but their scale is different from that of the distribution-function curves. The unimodality when $c > 1$ and absence of a mode when $c \leq 1$, alluded to in Section 2.2, is here seen for fits to real data.

The conditional response rate (CRR), which formed the theoretical basis for deriving the Weibull distribution in Section 2.1, is here given for real data, that for Item 16 again. Curve 1 in Figure 15 shows the CRR for the OK subgroup, with $c = 1.16$, and Curve 2 that for the NO subgroup with $c = .38$. Curve 1 increases monotonically with time, indicating (loosely) that the longer a person sticks with Item 16 the more likely it becomes that he/she will get it right if indeed he/she gets it right at all. Curve 2, on the other hand, decreases rapidly with time. Among people who do not get Item 16 right, the longer they stick with it, the less likely it becomes that they will respond to it the next instant, given that they haven't responded to it so far. In other words, many subjects gave the wrong answer early on but the "giving up" rate slows down as time goes by. It might be said that CRR expresses the degree of <u>involvement</u> in an item by examinees. But further explication of this concept must await further research.

---

[4] In fact, it was this relationship with the speed of "<u>convergence</u>" to asymptote that led us to denote the shape parameter as c.

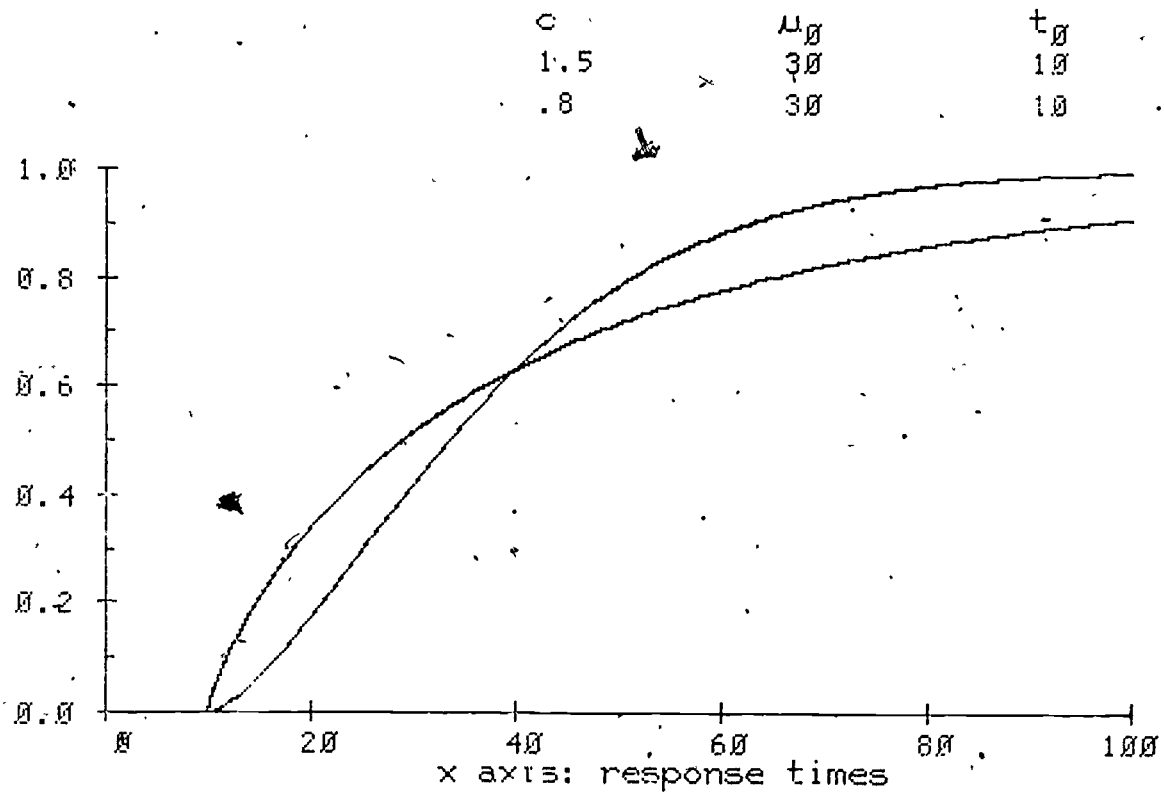
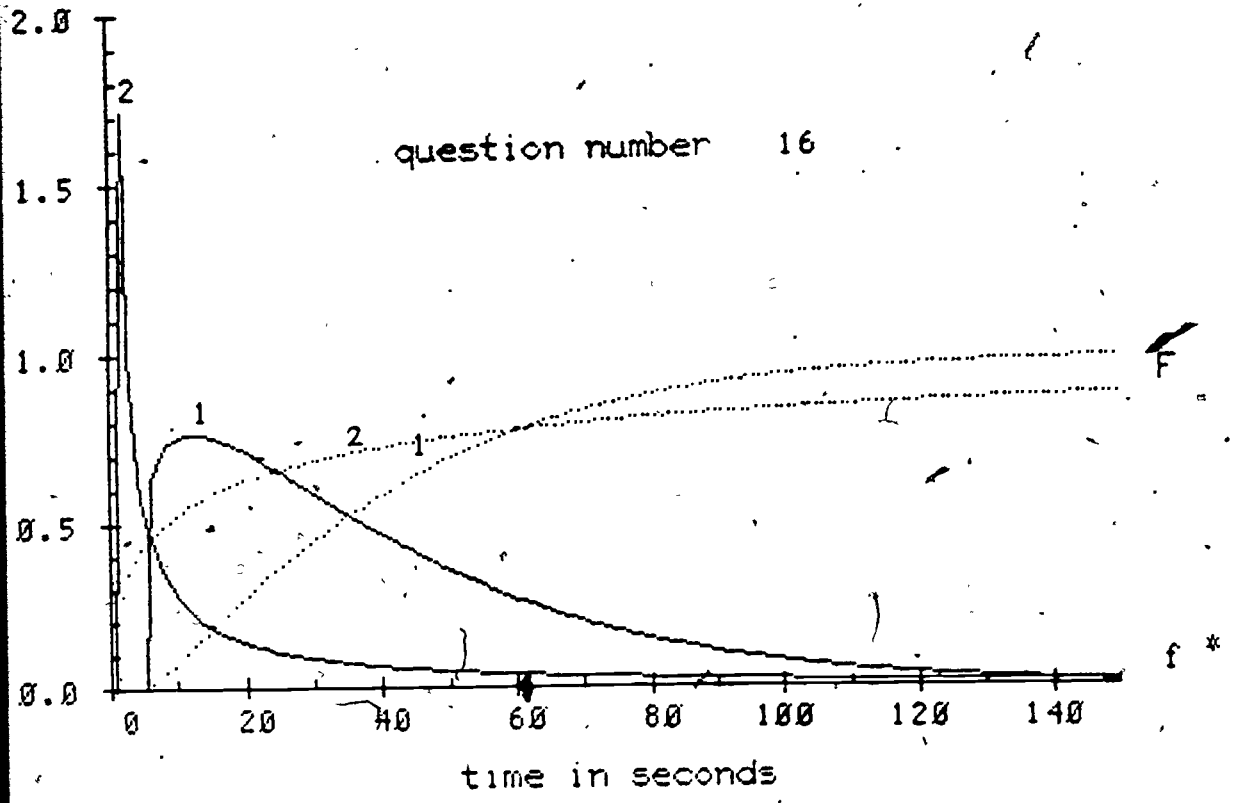


Figure 13 Weibull distributions for c=.8 and 1.5.

	c	$t_{\emptyset}$	$\mu_{\emptyset}$	mean
= okonly	1.161	5.4740	38.47	36.51
= noonly	Ø.3764	Ø.9206	19.33	76.68



* vertical scale for f(t) is magnified by a factor of $\mu_{\emptyset}$

Figure 14 Weibull distribution and density function

	c	t_0	μ_0
1 = okonly	1.161	5.4740	38.47
2 = noonly	0.3764	0.9206	19.33

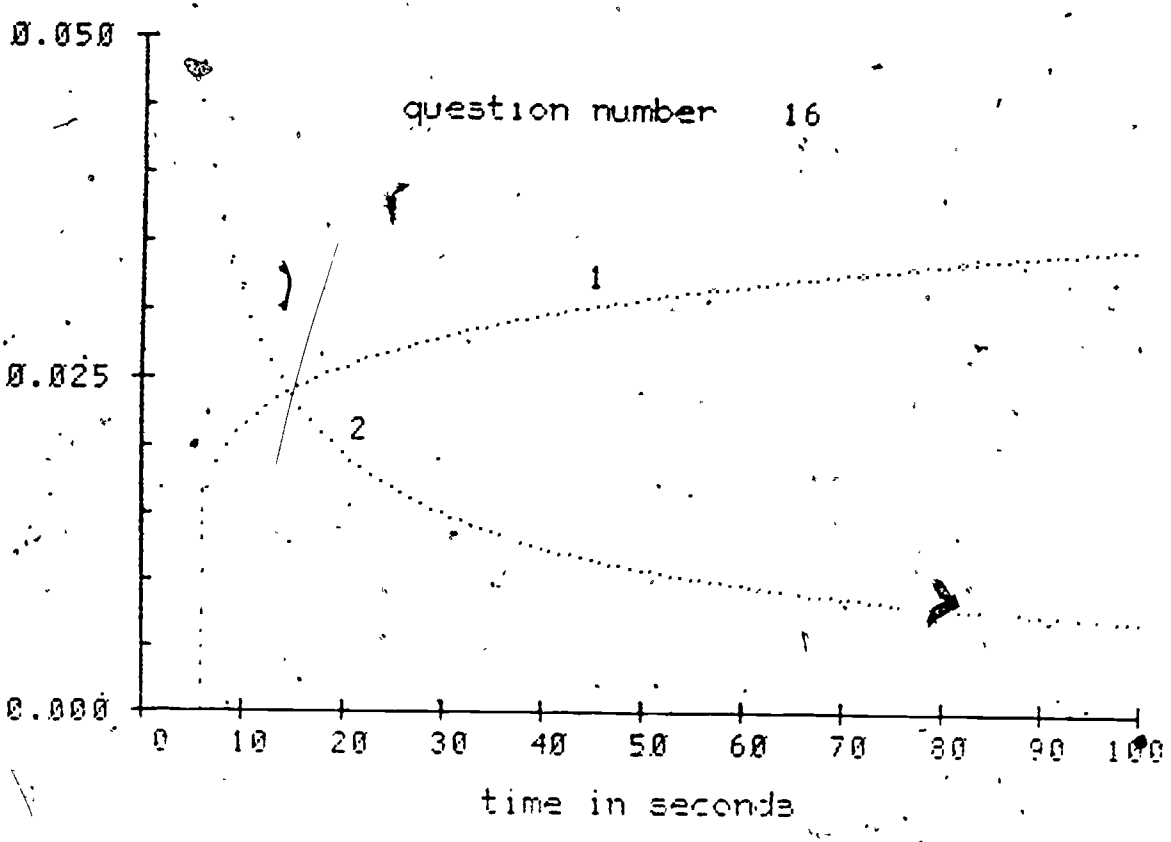


Figure 15 Conditional response rate function F(t)/(1-F(t))

Next, we identified five sets (four pairs and one triplet) of items that respectively had the same difficulty index in the traditional sense and were on the difficult side. The c-values (determined for the OK group) of the items within the "isodifficulty" sets were consistently and fairly substantially different. Examination of the item contents revealed that the c-values seemed to reflect a more intuitively plausible notion of "item difficulty" than did the traditional difficulty index. The data are as follows:

Difficulty Index (Proportion Passing)	Item Number	c-value (OK group)	Item Content
.279	38	1.062	A property of orthogonal transformations
.279	43	.982	Variances and eigenvalues
.314	12	1.336	Tricky problem on order of matrices
.314	37	1.019	Property of orthogonal transformations
.290	17	1.173	Symbols for matrix/vector operations
.290	23	1.014	If AB = AC then B = C
.290	33	1.240	Matrix inverse: numerical example
.349	35	1.917	Orthogonal transformation: numerical example
.349	39	1.276	Simple property of orthogonal transformations
.442	21	1.126	Row-wise expansion of determinants
.442	36	.784	Property of orthogonal transformations

The foregoing data suggest that the Weibull parameter c may be a more sensitive measure of the conceptual difficulty of an item than is the traditional difficulty index defined as the proportion of examinees getting the item right. In fact, for the OK group the items are completely undifferentiable by the traditional difficulty index, since the value is 1.00 for every item. Yet c enables us to differentiate among such items by detecting different rates at which the asymptote is approached.

Items 36 through 39 all ask the simple properties of orthogonal transformations. Their difficulty indices are .442, .314, .279 and .349, respectively, but their c values increase monotonically in the order the items were presented: .784, 1.019, 1.062, 1.276. This makes

sense when we consider the concept of, or the reasoning behind, the Weibull parameter c in general. The CRR for questions related to the same topic seems to increase as the familiarity with the topic increases, as it should from earlier to later items on the same topic. Thus, the parameter c seems to be related to what may be termed degree of involvement on the one hand and degree of familiarity on the other. Both these are indirectly related to difficulty but are conceptually different from it.

To conclude, the means, across the 48 items, of the three Weibull parameters in the two subgroups were as follows:

	t_o (secs)	c	μ_o
OK subgroup	2.7	1.125	33.05
NO subgroup	1.1	.903	22.50

We did not discuss the scale parameter μ_o in the foregoing, but in view of its mathematical relation, $\mu = t_o + \mu_o \Gamma(1+1/c)$, with the theoretical mean of the distribution, it hardly needs discussion. Since the mean of μ_o is smaller in the NO subgroup than in the OK subgroup, we may conclude that, on the average, the NO subgroup spent less time per item than did the OK subgroup in the matrix algebra pretest.

5.5 Correlations among Weibull Parameters and Item Statistics

The foregoing concludes the main analyses carried out on the data from the original version of the matrix algebra pretest. To explore other possible relations, however, the three Weibull parameters and the maximum correlation between $\ell n \ell n (1-P)^{-1}$ and $\ell n (t-t_o)$ that was found in the process of estimating the parameters (see Section 3.1) for the OK subgroup and the NO subgroup separately were correlated with the Kolmogorov-Smirnov p-values in the OK subgroup and six other item statistics based on the total sample. There were thus 15 variables in all, but the μ_o in the NO subgroup had to be omitted because of storage limitations. The resulting 14 x 14 correlation matrix is shown in Table 10, where the correlations significant at the 5 percent level are asterisked. (All correlation coefficients used as input variables were transformed into Fisher's Z before being correlated with other variables.)

Table 10

A Correlation Matrix of Weibull Parameters and Item Statistics

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. t_o (OK sub-group)	1.000													
2. max. correlation	.067	1.000												
3. c	-.169	.185	1.000											
4. μ_o	-.285*	-.184	.245	1.000										
5. No. of options	-.066	-.179	.316*	.488*	1.000									
6. p from Kolmo.	.095	.143	-.331*	.033	.023	1.000								
7. Difficulty 1	-.007	.155	.406*	-.137	.113	-.611*	1.000							
8. $r_{s,i}$	-.029	.109	.160	.100	-.049	-.216	.241	1.000						
9. $r_{s,ti}$	-.210	.124	.097	.340*	-.066	.271	-.492*	-.018	1.000					
10. Difficulty 2	.016	.131	.374*	-.169	.163	-.582*	.991*	.175	-.544*	1.000				
11. Average time	.131	-.149	.183	.912*	.541*	-.015	-.004	.089	.219	-.030	1.000			
12. t_o (NO sub-group)	.016	-.048	-.003	.143	.080	-.167	.187	.009	-.189	.188	.259	1.000		
13. max. correlation	.088	.000	.091	.117	-.108	.140	-.243	-.021	.222	-.262	.083	-.004	1.000	
14. c	.240	-.031	-.170	-.189	-.221	-.024	-.153	-.232	-.133	-.136	-.157	-.108	.186	1.000

*Significant at $p < .05$; i.e., $r = -.285$. N = 48 items.

Note. All correlations were converted by Fisher's Z-transformation. The first 4 variables are of OK subgroup, the last 3 variables are of NO subgroup.

The Weibull parameter c of the OK subgroup correlates with the numbers of options in the items (.316), difficulty 1 (.406), the arcsine transform of the Kolmogorov-Smirnov p-values (-.331), and difficulty 2 (.374). The negative correlation with the Kolmogorov-Smirnov p-value is probably largely an artifact, because items with small values of c tend to be more difficult and hence the OK subgroup for those items tend to be smaller, thus inflating the "significance" and decreasing the p-values. When the difficulty index is partialled out, the correlation between c and p drops to -.115, which is nonsignificant.

The shape parameter in the OK group also correlates .316 with the number of options, meaning that items with more choice options (which ranged from two for true-false items to five for the multiple-choice item with the largest number of alternatives) tended to have larger c values. If our interpretation, suggested earlier, that the item c-value reflects degree of engagement students show with the item is correct, we may conclude that within the range represented, the larger the number of options the greater the engagement students feel. This seems reasonable since items with more alternatives present more of a cognitive task and hence probably induce greater involvement on the part of students. To put it the other way around, this observation lends further support to our notion that c reflects degree of engagement. It should be mentioned that partialling out Variable 7 (difficulty 1) does not affect the correlation (the partial r is .30); hence the correlation between c and number of options cannot be explained away by arguing that the larger the number of options the more difficult the item tends to be.

It is interesting, although rather disappointing, to note that none of the Weibull parameters correlates with item discriminating power, $r_{s,i}$.

Next, the scale factor μ_o correlates moderately to highly with the number of options (.488), $r_{s,ti}$ (.340) and the observed average time (.912). Since μ_o is functionally related to μ (the theoretical observed time--see equation (2.7)) its very high correlation with observed average time is also to be expected. This in turn explains the moderate correlation between μ_o and number of options, since the larger the latter is the more time it would take, by and large, to respond to that item.

The p-values from the Kolmogorov-Smirnov tests of goodness-of-fit in the OK subgroup have correlations -.611 and -.582 with difficulty 1 and difficulty 2, respectively. This, too, is probably an artifact to a large extent in that the p-values tend to decrease as sample size

increases, and large n for the OK subgroup means an easy item whose "difficulty" indices would be large.

The correlation $r_{s,ti}$ between item response time and total test score in the total sample of 88 subjects correlates with μ_o (.340), difficulty 1 (-.492) and difficulty 2 (-.544). The negative correlation with the difficulty index has already been discussed in the previous subsection. The positive correlation with the scale factor μ_o is difficult to interpret.

It should be noted that the Weibull parameters, t_o, r_{max} and c, from the NO subgroup did not correlate significantly with any of the other 11 variables. This is probably because the data analyzed here is for the pretest, and members of the NO subgroup know little if any matrix algebra. The situation changes considerably when we analyze the posttest data in the next section.

6. ANALYSIS OF POST-REVISION DATA

After the results from the original version of the matrix algebra pretest were analyzed, and partly as a consequence of the analyses, several items were modified to correct ambiguities in wording or defects in the display. Another change made in the test was, as mentioned earlier, that the option of pressing the NEXT key to go to the next item without answering the previous one was eliminated. This change was made at the request of the instructor of one of the participating statistics courses who wanted to force the students to answer all questions. In retrospect, however, this may have been a change for the worse, for it has no doubt led to increased guessing. At the risk of seeming to attribute to the Weibull distribution some magical power to detect "undesirable" items, we note that the fit became very poor for items in which a large increase of guessing must have occurred such as those testing for difficult material like transformations. It could be that guessing contaminates the distribution so that it no longer appears to result from a single underlying stochastic process. How to handle this problem is something we are not prepared to say at this time. That must be left to future research.

In this section we discuss the analyses not only of data from the revised pretest but those from the posttest as well. (In fact, the analyses are mostly of the posttest data.) We must therefore first describe those tests.

6.1 Description of Posttests (with Some Speculations)

It might have seemed strange that the analyses discussed so far were confined to pretest data, with no mention of a posttest. This was simply because no posttest had existed before Fall 1976. Thanks to NIE funding, we were able to implement four posttests during that semester. (Lest it be thought that funds were diverted from research to instructional use--especially in the current atmosphere of censure of mishandling of grant monies!--it should be pointed out that the posttest results were not used at all in determining grades in the three statistics courses. Thus, these tests served our research purposes only.)

Specifically, the tests come after completion of the lessons on matrix multiplication, on matrix inversion, on transformations, and on eigenvalue problems, and they are referred to as "Multpost," "Matinvtest," "Transtest," and "Eigtest," respectively. The items on each constitute a subset of the 48 items on the pretest, there being 23 items in Multpost, 12 in Matinvtest, 7 in Transtest, and 8 in Eigtest. (The numbers total 50 because the first two tests contain two items in common.) The actual items, identified by which numbers they are in the pretest, are shown on the following page along with the number of subjects.

on whom usable data for each test were available. (The interested reader may refer to Appendix B to see what the items on each posttest are.)

	Items	Number of Subjects
Multpost	1-18; 25-29	68
Matinvtest	17-24; 30-33	30
Transtest	34-40	38
Eigtest	41-48	56

We mention in passing that many students complained, in their responses to an open-ended question included in the questionnaire attached to the lesson, that some of the items in Transtest tested for material beyond what was taught in the lesson on transformations. They claimed that these items were too mathematical and advanced for the students to whom the test was addressed. Despite these complaints, however, we did not modify these items because we were curious to see whether the Weibull fitting would be adversely affected by the lesson-unrelatedness of the items.

Three Items with Matrices Constructed by a Random Number Generator. In three of the items, the elements of the matrix in the stem were generated by a random number generator, so students would not get identical matrices to work with. Each had a parallel, counterpart item in which the elements of the matrix were fixed. We were curious to see whether the two types of item would lead to equal degrees of fit to the Weibull. If not, it would indicate that the time taken for the sheer arithmetical calculations, which may vary from version to version of the randomly generated items, plays an important part in the total time taken for the item, thus leading to a more complicated distribution with separate Weibull processes for the arithmetic and the matrix algebra parts. A twofold Weibull convolution might then offer a better fit to the "random" items while a regular Weibull would fit the "fixed" items. Alternatively, if the component parts are successfully modeled by one- (two-)parameter negative exponential distributions, then both the fixed and random items would be well fitted by two- (three-)parameter gamma distributions with the random items having a value for c greater by approximately 1 than the fixed items.

Posttests Should Make Students More Seriously Involved in Solving Items. Since the conditions under which pretests and posttests are taken differ considerably, we expect that the incidence of guessing will differ in the two cases. Specifically, we expect that guessing will be minimized in the posttests while the press to guess would be greater in the pretest, especially when the option to skip items is eliminated. Also, the knowledge of matrix algebra newly acquired after the students have gone through the lessons will have led them

to be more involved with answering the test questions. Consequently, the CRR should be monotinically increasing with time rather than decreasing or remaining constant. We would therefore expect c to be greater than 1 for posttest items. On the other hand, if the student doesn't know or has forgotten the material, we would expect him/her more likely to give answers by random guessing.

6.2 Results of Analyses

Tables showing p-values and z-values from the Kolmogorov-Smirnov tests of goodness-of-fit and the Weibull-parameter values for the items in the four posttests are given in Appendix E. Here we discuss only the summary results and their implications. Table 11 shows the percentages of items having Kolmogorov-Smirnov p-values greater than .20, greater than .40 and greater than .50 and the mean p-values for the four posttests in the two subgroups. Also shown for comparative purposes are the mean p-values in the pretest.

Table 11

Percentages of Items with Kolmogorov-Smirnov p-values Exceeding Three Values in the Two Subgroups

	p > .20	p > .40	p > .50	$\bar{p}$	Pretest $\bar{p}$
OK Subgroup					
Multpost	87%	83%	78%	.65	.46
Matinvtest	100%	92%	75%	.64	.55
Transtest	71%	57%	43%	.50	.51
Eigtest	100%	100%	100%	.79	.78
NO Subgroup					
Multpost	100%	100%	100%	.82	.51
Matinvtest*	----	----	----	---	.50
Transtest	71%	71%	57%	.54	.43
Eigtest	100%	100%	88%	.80	.27

*Insufficient data

The distributions of time-score data from all but the Transtest fitted the Weibull distribution much better than did those of the revised pretest, for which only 79 percent of the items had p-values

in excess of .20 in the OK subgroup and 75 percent in the NO subgroup. The items in the pretest corresponding to those in the Transtest had an average p-value of .51 in the OK subgroup and .43 in the NO subgroup, as shown in Table 11. The average p-values in the Transtest in the two groups, on the other hand, were .50 and .54. As we mentioned earlier, there were some items in the Transtest that covered material not taught in the lesson, inviting much student complaint. Again, something unusual in the items seems to result in poorer Weibull fit. Apart from this, the mismatch between the lesson and the test should play a role in the study of the importance of the linkage between lesson and items in measuring the effectiveness of instruction as well as in assessing the level of a student's learning. Such a study is planned in a forthcoming project.

Let us examine in some detail the results for the problematic test, Transtest, which included items covering material not taught in the lesson. (It may be mentioned in passing that some students expressed irritation and hostility, while others thought they had missed something in the lesson and went back to repeat it.) For comparative purposes, the averages of the Weibull parameters t_o and c for items, in three posttests, with adequate numbers of subjects in both the OK and the NO subgroups are shown below.

		t_o	c	
Multpost	OK subgroup	4.98	1.11	(based on 8 out
	NO subgroup	9.45	1.06	of 12 items for NO)
Transtest	OK subgroup	6.83	1.04	(7 items)
	NO subgroup	5.56	.88	
Eigtest	OK subgroup	7.84	1.22	(8 items)
	NO subgroup	6.09	1.28	

In brief, the Transtest averages alone out of the three posttest averages shown above exhibit a pattern typical of that for a pretest item, which was exemplified by Item 16 in the previous section (see page 44). That is, the NO subgroup has smaller values for both t_o and c than does the OK subgroup. In particular, the c value in the NO subgroup is smaller than 1 while that in the OK subgroup is greater than 1. (Note that in neither of the other two posttests do the average t_o and the average c show all of these relations between OK and NO subgroups.) Thus we may conclude that the Transtest, although a posttest, acted much like a pretest by virtue of the anomalous items. Again, there is corroboration of our speculation that c indicates extent of engagement or involvement on the part of students in an item. The fact that the average c for Transtest in the NO subgroup is only .88 suggests that many students merely guessed at the

answers for the three items in this test that covered material beyond the scope of the lesson, which is only to be expected.

As mentioned earlier three items used matrices whose elements were chosen by a random number generator supplying integers between -9 and 9 inclusive. Specifically, Item 1 asked for the sum of two 3 x 3 matrices with fixed elements; Item 2 was a parallel item with random elements. Items 3 and 4 called for the difference between two 3 x 3 matrices with fixed and random elements, respectively. Items 5 and 6 asked for the transpose of a 3 x 3 matrix, again with fixed and random elements respectively. That is, the odd-numbered items used fixed matrices while the even-numbered items used random matrices.

Table 12 shows the Kolmogorov-Smirnov p-values and the Weibull parameters c and μ_o for these three pairs of items in the revised pretest and in the posttest for the OK and NO subgroups (except that the latter subgroup is nonexistent for the posttest because everyone got all six items correct). Note that for the OK subgroup the even-numbered items have considerably larger c values than do their odd-numbered counterparts in the revised pretest, while the reverse is true for the NO subgroup. Interpretations will be attempted after the findings have been stated factually.

Table 12

Comparison of Items Using Fixed Matrices (Odd-numbered) and Parallel Items (Even-numbered) Using Random Matrices in Terms of Kolmogorov-Smirnov p-values and the Weibull Parameters c and μ_o

	OK Subgroup						NO Subgroup		
	Pretest			Posttest			Pretest		
Item	p	c	μ_o	p	c	μ_o	p	c	μ_o
1	.06	1.15	44.5	.33	1.06	25.3	.99	2.08	52.6
2	.02	1.92	28.1	.26	2.76	23.6	.99	.49	32.1
3	.64	1.24	23.3	.94	1.24	17.7	.41	1.05	19.3
4	.51	2.19	32.4	.17	1.05	20.9	.66	1.02	14.5
5	.45	1.25	13.7	.38	1.00	10.9	.99	2.10	17.1
6	.16	1.35	12.7	.08	1.41	11.6	.42	.85	10.8

Next the conditional response rates (CRR) of Items 1 and 2 are compared in Figure 16 for the OK subgroup and in Figure 17 for the NO subgroup. In the OK subgroup it is Item 2, which uses the random matrix, that has a larger and monotonically increasing CRR while the CRR for Item 1 is smaller and almost parallel to the horizontal axis. In other words the conditional probability (given that the item hasn't been answered up to then) that it will be answered the next moment is always greater for Item 2 than it is for Item 1. Note that, on the posttest, both items were answered correctly by all subjects and hence traditional item analysis based on the performance score would fail to show any difference between these two items that are identical in framework but differ in the way the matrix elements were chosen. Our analysis based on time scores has revealed an interesting difference as the c-values in Table 12 shows (although the Items 3 and 4 pair is an exception).

The CRRs of Items 2, 4 and 6 were larger than those of Items 1, 3 and 5, respectively, in the OK subgroup on the revised pretest, but the reverse was true in the NO subgroup. That is to say, the conditional probability that Item 1 will be answered wrong in the next moment increased with time while the same conditional probability decreased for Item 2. Similar remarks hold for the Items 3 and 4 and Items 5 and 6 pairs. Figures for these pairs in both subgroups are shown in Appendix F.

Interpretations. The complicated findings reported above are difficult to interpret, and what follows must be regarded as attempts rather than definitive interpretations. The three pairs of items were very simple for students taking advanced statistics courses in education and psychology once they learned the definitions of matrix addition, subtraction and transposition. The only difficulty probably occurred in the addition and (more so) the subtraction of signed numbers, and in the case of transposition, the requirement to rapidly perceive and distinguish among five 3 x 3 matrices (the options) with the same set of numbers in different arrangements. We surmise that the items with fixed matrices (Items 1, 3 and 5) were so easy for the OK subgroup that the answers were arrived at without much "involvement" on the part of the students. But some versions of Items 2, 4 and 6 probably required greater attention and involvement of the students, depending on particular combinations of numbers chosen by the random number generator.

Let us now return to Table 12 and examine the Weibull parameters for the posttest. In item pairs 1, 2 and 5, 6 they followed exactly the same pattern as they did for the pretest. In fact, the difference in c values between the odd- and even-numbered members of the pairs were greater for the posttest than they were for the pretests. Referring to equation (2.4) for the Weibull CRR function, it can be seen that for fixed c, this is a monotonically decreasing function of μ_o. Since all μ_o's for the posttest were smaller than the corresponding μ_o's for the pretest, the CRR values at any given

		c	t_0	μ_0
item	1 :	2.0879	0.0001	52.64
item	2 :	0.4865	11.4871	32.09

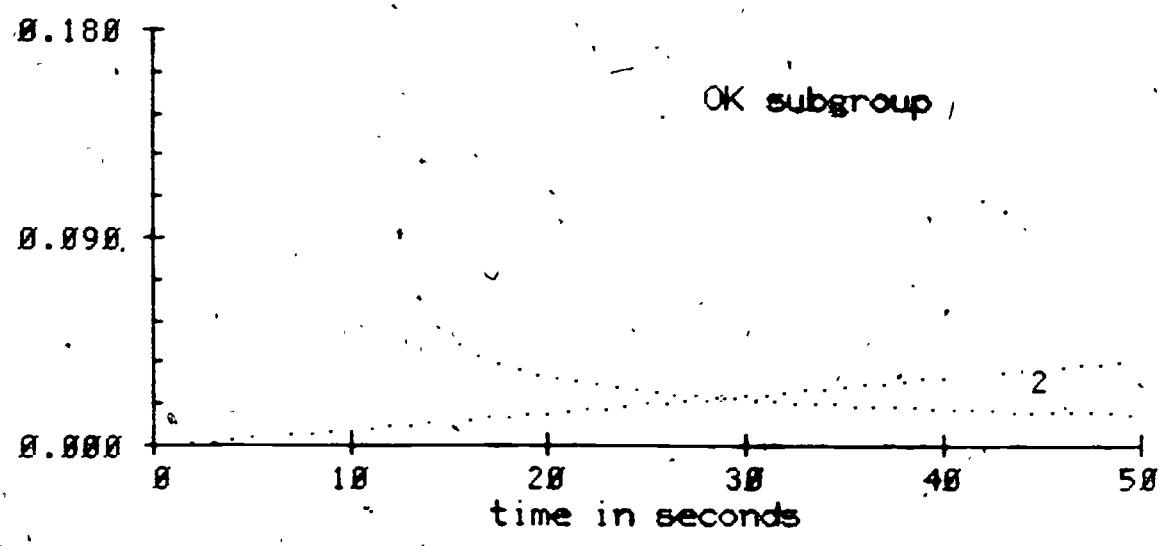


Figure 16 Comparison of conditional response rates of items 1 and 2 for OK subgroup

		c	t_0	μ_0
item	1 :	1.150	10.5903	44.51
item	2 :	1.924	2.5260	28.12

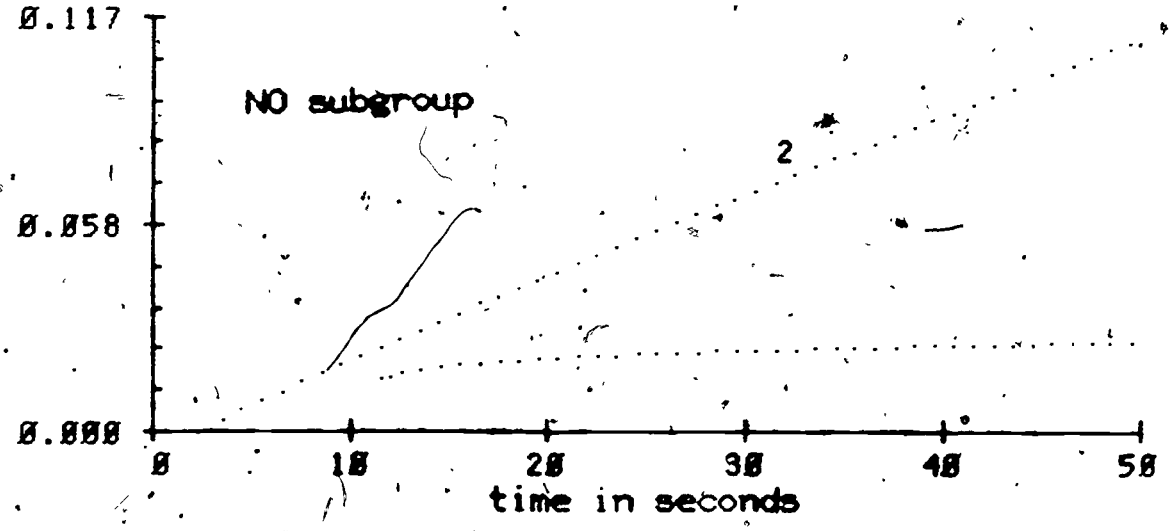


Figure 17 Comparison of conditional response rates of items 1 and 2 for NO subgroup

60a

time point were mostly larger for the posttest than they were for the pretest. In item pair 3, 4 the pattern was not the same for the posttest as it was for the pretest.

For members of the NO subgroup, the materials of these three item pairs were relatively new and unknown or were forgotten. As time went by the CRR's slowly increased for the fixed-number items 1, 3 and 5, while this was not the case for Items 2, 4 and 6. (See figures in Appendix F.) This means that the students who answered the fixed-number items wrong tried harder, on the average, to figure out the right answers than did the students who answered the random-number items. Two of the latter items (Nos. 2 and 6) were given up quickly on the average, again suggesting that some number combinations chosen by the random-number generator led to difficulties. In the exceptional pair, Items 3 and 4, the c values were almost equal and close to one (1.05 and 1.02) but the μ_0's were somewhat different (19.3 and 14.5). Therefore the CRR curve for Item 4 lies above that for Item 3, and both are almost horizontal. Recalling that Item 4 differs from Item 2 only in that the operation is subtraction instead of addition, we infer that the difficulty of subtraction of signed numbers depends strongly on the particular pair of numbers involved. This must be the reason why the item pair 3, 4 showed a different behavior from item pairs 1, 2 and 5, 6 in intrapair differences.

7. WEIBULL AND GAMMA FITS COMPARED

In this section we compare the relative goodness of fit of the Weibull and two-parameter gamma distributions to time-score data from various sources and attempt to come up with an explanation for when and why which offers the better fit. As a general rule, it seems that for material requiring a sustained, uniform thinking process the Weibull has an edge over the gamma. The Weibull also shows wider applicability and greater flexibility. On the other hand, if the task consists of a concatenation of several relatively independent and more or less simple, mechanical subtasks or stages, the gamma fit seems better. Of course pure cases of either type are rare, and often the fits are ambivalent. Cases for which neither distribution offers an adequate fit seem to be ones in which the several stages of a task are non-independent or non-mechanical or both. Compound Weibull distributions would probably show good fits in such cases.

Due to the abundance of fittings undertaken, the Kolmogorov-Smirnov p-values and the Weibull and gamma parameter estimates are shown in Appendix E, and only summary tables are given in this section.

7.1 Multiplication Pretest and Posttest

Data from a sample of 56 examinees who took both the pretest and posttest for simple operations and matrix multiplication lessons were subjected to goodness of fit testing for Weibull and gamma distributions. The estimated parameters of Weibull and gamma distributions based on time-score data of 23 items were determined and tested by the Kolmogorov-Smirnov test. Items were classified into four categories according to p-values from the Kolmogorov-Smirnov testing:

(1) p values for Weibull (p_W) is much better than p values for gamma (p_G); ($p_W - p_G \geq .10$)

(2) p_W is better but not much; ($.10 > p_W - p_G > 0$)

(3) p_G is better but not much; ($.10 > p_G - p_W > 0$)

(4) p_G is much better than p_W; ($p_G - p_W \geq .10$).

In order to show which theoretical distribution is better for the 23 items, the frequencies in each category were counted and summarized in Table 13.

Table 13

Comparison of Goodness of Fit for Weibull and Gamma in Multpost

	Pretest (23 items)		Posttest (23 items)	
Category	OK Subgroup	NO Subgroup*	OK Subgroup	NO Subgroup**
1	13 (57%)	9 (47%)	16 (70%)	6 (75%)
2	4 (17%)	4 (21%)	2 (9%)	1 (12.5%)
3	3 (13%)	6 (32%)	3 (13%)	0
4	3 (13%)	0	2 (9%)	1 (12.5%)

*Four items were omitted because of small N.

**Only eight items had $N \geq 12$.

The Weibull distribution is a distinct preference for both the OK and NO subgroups in the pretest and the posttest, but the posttest shows a slightly higher percentage in Category 1 than the pretest does. The items that fall in Category 4 are only a few in each group

of the pretest and posttest. It can be said of 87 percent, 100 percent, and 92 percent of items in both groups of the pretest and OK subgroup of the posttest that the cumulative distributions of their time score data are pretty well approximated by Weibull distribution functions.

The average p values and standard deviations of each group for both distribution functions are given in Table 14.

Table 14

Average p-values for Weibull and Gamma Distributions; Multpost (23 items and N = 56)

	Pretest		Posttest	
	OK Subgroup	NO Subgroup	OK Subgroup	NO Subgroup
$\bar{p}_W$	.73	.75	.65	.82
SD_W	.28	.30	.27	.16
$\bar{p}_G$	.59	.63	.39	.50
SD_W	.37	.35	.37	.35

The average values of p_W in the four columns are larger than those of p_G in the same columns and the standard deviations of p values for Weibull (SD_W) are smaller than the standard deviations of p for gamma (SD_G). The values of p_G fluctuate considerably more than those of p_W. The result that Weibull is better was expected because items in the matrix algebra test were not easy, and many levels of knowledge that are hierarchically or linearly related would have been required to arrive at their responses. Therefore a gamma distribution, which is a convolution of finite number of independent negative exponential variables which Restle (1962) interpreted as representing independent stages or components of a problem solving process, cannot explain theoretically the time score data from matrix algebra test items where the stages or components are not independent. Indeed, for most of our items there is no way we can say that the stages to reach a response of a given item are independent from one another. Since the Weibull distribution does not require such a strong assumption so, no matter how each stage relates one to the other, a whole process of cognitive tasks to reach a response can be modeled by a Weibull distribution. The response can be positive or negative as long as a student's process of achieving their cognitive task can be considered to be of the same kind. This

means that the OK subgroup in the posttest and NO subgroup in the pretest follow different processes of thinking for reaching their responses. but subjects in the OK subgroup may be following a very similar thinking process to reach their responses, and so are subjects in the NO subgroup for the pretest.

The NO subgroup in the pretest may be characterized as follows: many examinees gave up trying a problem hard and responded by guessing their answer while the students in the NO subgroup in the posttest tried hard and spent longer times but unfortunately their answers were wrong. A close examination of the CRR would tell more about these relations.

It is interesting to note that the p_G of the pretest was .59 but it dropped to .39 for the posttest in Table 14, and for the NO subgroup, it dropped from .63 to .50 while $\bar{p}_W$ of both groups don't change their values so much. As we mentioned earlier in this paper, if responses to a given item occur at a random base, then the time-score data follows a negative exponential function. Gamma is a convolution of such negative exponential functions. It is probably true that the number of examinees who took the pretest answering randomly by guessing are likely larger than those for the posttest by which time everybody had learned the material already.

The Revised Pretest, the Case of N = 100 and 48 Items. Although the original version of the pretest was designed so as to minimize the guessing effect on the time score data, their data were not analyzed for comparative study of goodness of fit testing of Weibull and gamma distributions. The revised version of the pretest has a matched sample of Multpost as a subset of the N = 100, the whole pretest sample, for items 1-18 and 25-29 out of the 48 items and these 23 items were analyzed in the previous subsection, but the summary of the pretest, 48 items is given below.

Table 15

Comparison of Weibull and Gamma Fitting for the Revised Pretest (48 items and N = 100)

Category*	OK Subgroup	NO Subgroup
1	29 (60%)	22 (46%)
2	6 (13%)	11 (23%)
3	11 (23%)	12 (25%)
4	2 (4%)	3 (6%)

*1 = {items with $p_W - p_G \geq .10$};
2 = {items with $0 < p_W - p_G < .10$};
3 = {items with $0 < p_G - p_W < .10$};
4 = {items with $p_G - p_W \geq .10$}

In Table 15, only two items for the OK subgroup and three items for the NO subgroup fell in Category 4; that is, 94 percent to 96 percent of the 48 items are favorable to Weibull distributions for both the OK and NO subgroups. The average p values and standard deviations have very similar results to those of the 23 matched pretest and posttest items.

Table 16

Average p-values for Weibull and Gamma; the Revised Version of Pretest (48 items and N = 100)

Distribution	Mean S.D.	OK Subgroup	NO Subgroup
Weibull	$\bar{p}_W$	.59	.52
	SD_W	.33	.36
Gamma	$\bar{p}_G$	.47	.40
	SD_G	.38	.41

The average values of p_W and p_G for the OK and NO subgroups are about .10 smaller than the average values shown in Table 14 and the standard deviations in Table 16 are larger than those in Table 14. But it is obvious that our observation in the previous section is applicable to these data as well.

The Posttest: Matinvtest, Transtest, and Eigtest. Three more posttests were analyzed by the Kolmogorov-Smirnov test. These samples were not matched with the pretest sample. The results of a close examination for these data only revealed the same conclusion as those in the previous two subsections, with more emphasis on the fact that Weibull distributions are more suitable to our items of the posttests in the matrix algebra test than gamma distributions. Tables 17 and 18 summarize our observations.

We have only two items which fall in Category 4 in Table 17, one each in Transtest and Eigtest. It is natural to wonder which items fall in Category 4 and why their time scores fit gamma better. We will pick up such items and discuss further details in the following subsection.

Table 17

Comparison of Weibull and Gamma Fitting;
Matinvtest, Transtest and Eigtest

	Matinvtest*	Transtest		Eigtest	
	12 Items	7 Items		8 Items	
Category	OK Subgroup	OK Subgroup	NO Subgroup	OK Subgroup	NO Subgroup
1	9 (75%)	3 (43%)	4 (57%)	5 (62.5%)	2 (25%)
2	1 (8%)	2 (29%)	2 (29%)	1 (12.5%)	1 (12.5%)
3	2 (17%)	1 (14%)	1 (14%)	1 (12.5%)	5 (62.5%)
4	0	1 (14%)	0	1 (12.5%)	0

*Almost everybody got all 12 items correct, so the NO subgroup is almost empty.

Table 18

The Average p-values for Weibull and Gamma; Matinvtest, Transtest and Eigtest

		Matinvtest	Transtest		Eigtest	
Distribution	Mean S.D.	OK Subgroup*	OK Subgroup	NO Subgroup	OK Subgroup	NO Subgroup
Weibull	$\bar{P}_W$	.64	.50	.77	.79	.80
	SD_W	.21	.37	.34	.16	.21
Gamma	$\bar{P}_G$	.33	.42	.54	.65	.76
	SD_G	.38	.42	.34	.22	.21

*Almost everybody got all 12 items correct, so no analysis for the NO subgroup was carried out.

Items Whose Time Scores Fit Gamma Better. We found that Weibull distributions are generally more appropriate to approximate the cumulative distribution of the item time-score data from the matrix

algebra test than two parameter gamma distribution functions. But there are a few items whose p-values from goodness-of-fit testing are favorable to gamma. Of course, our sample size is not large enough and the observations are restricted to only 48 items in pretest and posttest, therefore it is dangerous to conclude that the Weibull definitely is our distribution. Besides, there are many psychological, intellectual and physical causes that individually or collectively may be responsible for reaching responses at any particular instant. It is impossible to isolate these causes and mathematically account for all of them, therefore the choice of response time distribution is still subjective and cannot be completely scientific. With these difficulties, it is necessary to appeal to a reasoning that makes it possible to distinguish between the different distributions on the bases of logical considerations.

We hope that our reasoning developed in the previous sections in terms of why most time-score data of items in the matrix algebra test are favorable to Weibull distributions is convincing to the readers. We must argue now why these few items show their favor to gamma. They are Items 2, 4, 6, 9 and 25 for the OK subgroup on the pretest, 18, and 42 for the same group on the posttest, and Items 18, 27, and 34 of the pretest, and Item 17 of the posttest for the NO subgroup.

Items Presented in the Posttest Without Proper Instructions. Items 17 and 18 fell in Category 4 when they occurred in Multpost but when they occurred in Matinvtest, both items went into Category 1. Because they are testing the knowledge of determinant that is not taught in the multiplication lesson, a great number of examinees in the multpost sample did not know about the determinant of a matrix. This complaint was confirmed by students' open ended questionnaire. Transtest invited the same complaint. Transtest items in Table 17 show more favor to Gamma than items in Matinvtest and Eigtest for the OK subgroups. Now, let us go back to 17 and 18. Since Items 17 and 18 and some items in Transtest are the only items that were given to the students prior to the related lessons being taught, the related topics would never be taught in a series of matrix algebra lessons, their responses might have been reached by different causes. Since these items were well fitted to Weibull in the pretest, psychological effects might be disturbing the determination of t_0, the minimum response time and CRR, conditional response rate or the shape parameter c.

Items 2, 4, 6 and 9. As mentioned before, we experimented with two types of items: one type used fixed numbers in each element of matrices, the other used a random number generator to fill in each element. Items 1, 3 and 5 ask for addition, subtraction and transpose of 3 x 3 matrices with fixed elements while Items 2, 4 and 6 ask for the same operations with randomly supplied integers between -9 and 9 inclusive. Item 9 asks for multiplication of a scalar to a matrix and the random number generator supplies integer elements between -9 and 9 except for zero. Comparison of Kolmogorov-Smirnov p-values is shown in Table 19.

Table 19

Comparison of p-values for Weibull and Gamma;
Items 1, 2, 3, 4, 5, 6 and 9, Matched Sample,
N = 56, OK Subgroup

Items	Pretest		Posttest	
	p_W, Weibull	p_G, Gamma	p_W, Weibull	p_G, Gamma
1	.69	.03	.63	.00
2	.27	.77	.73	.67
3	.89	.76	.93	.93
4	.99	1.00	.57	.46
5	.80	.12	.93	.08
6	.15	.27	.18	.20
9	.88	.92	.93	.76

Items 2, 4, 6 and 9 have higher p-values from Kolmogorov-Smirnov test for gamma than for Weibull in the pretest while Items 1, 3 and 5 have higher p-values for Weibull than for gamma in the pretest. But in the posttest, the p-values for Weibull became higher than or almost equal to those for gamma.

Weibull distributions are determined by three parameters, the minimum time t_o, a shape parameter c, and a scale parameter μ_o while our gamma distribution has only two parameters, without a location parameter t_o (or minimum time). Graphic display of both of the cumulative distribution of time-score data and the theoretical distribution function on the same PLATO screen often shows that the smooth curve gamma did not fit the cumulative distribution step function near the initial point t_o.

The examples shown in Figures 18 and 19 explain the situation intuitively. Two-parameter gamma distributions lack the capacity to provide information about the minimum time t_o, unlike a three-parameter gamma which has a location parameter.

Since Items 1, 3 and 5 have fixed number elements in the matrices, the degree of difficulty due to calculation for each item is constant, and does not vary from item to item, while Items 2, 4 and 6, having a different set of numbers as elements in 3 [illegible] matrices, lead to different difficulties in calculations. For example, 9-1 is much easier than -9-(-1), especially for those who are wondering how to do the subtraction of two matrices. Thus, t_o minimum time to respond to an item such as Items 2, 4, 6 and 9 can be different from item to item, depending on what kind of numbers were picked up by the

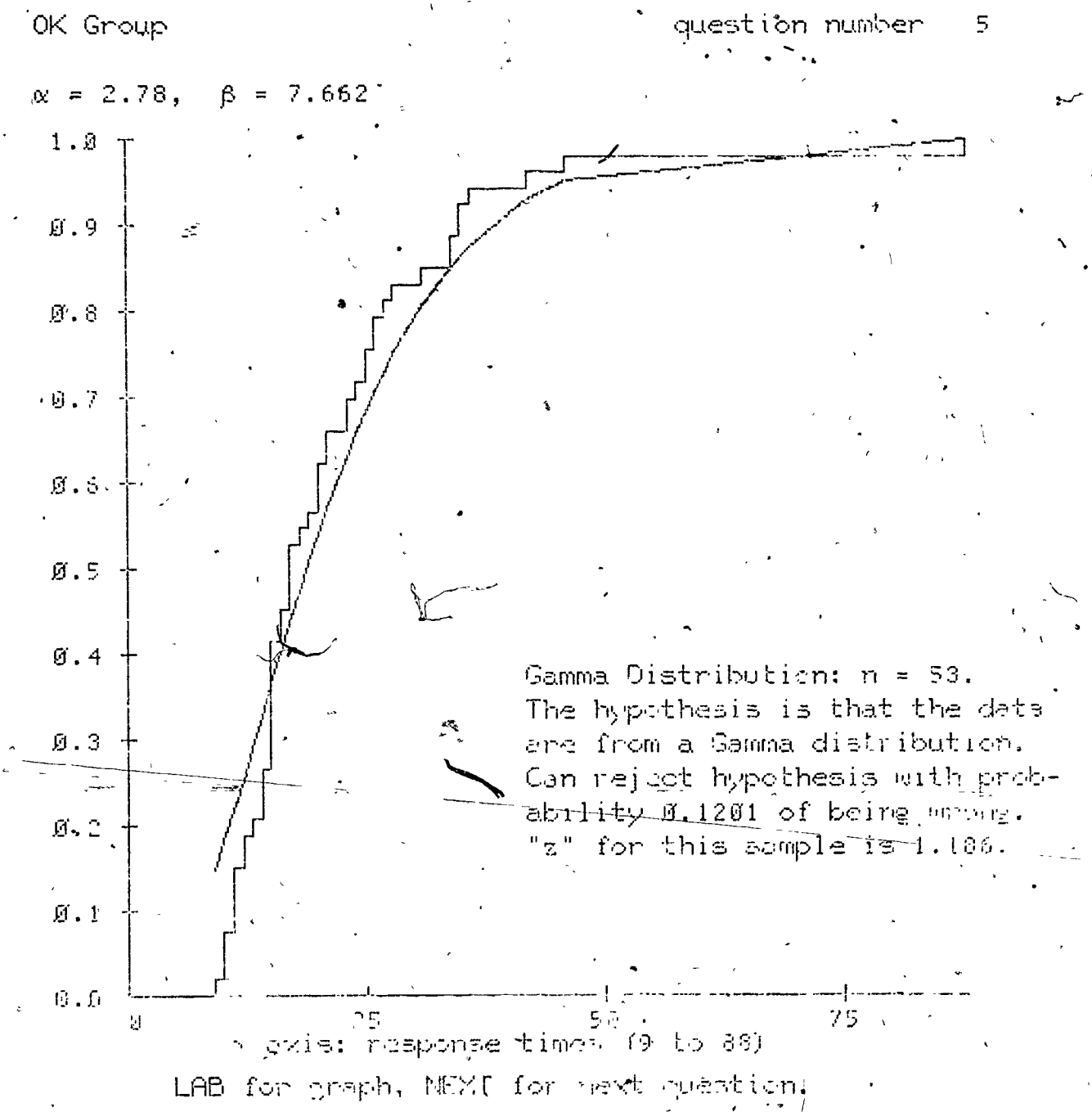


Figure 18 Goodness of fit test for the time-score data and Gamma distribution function

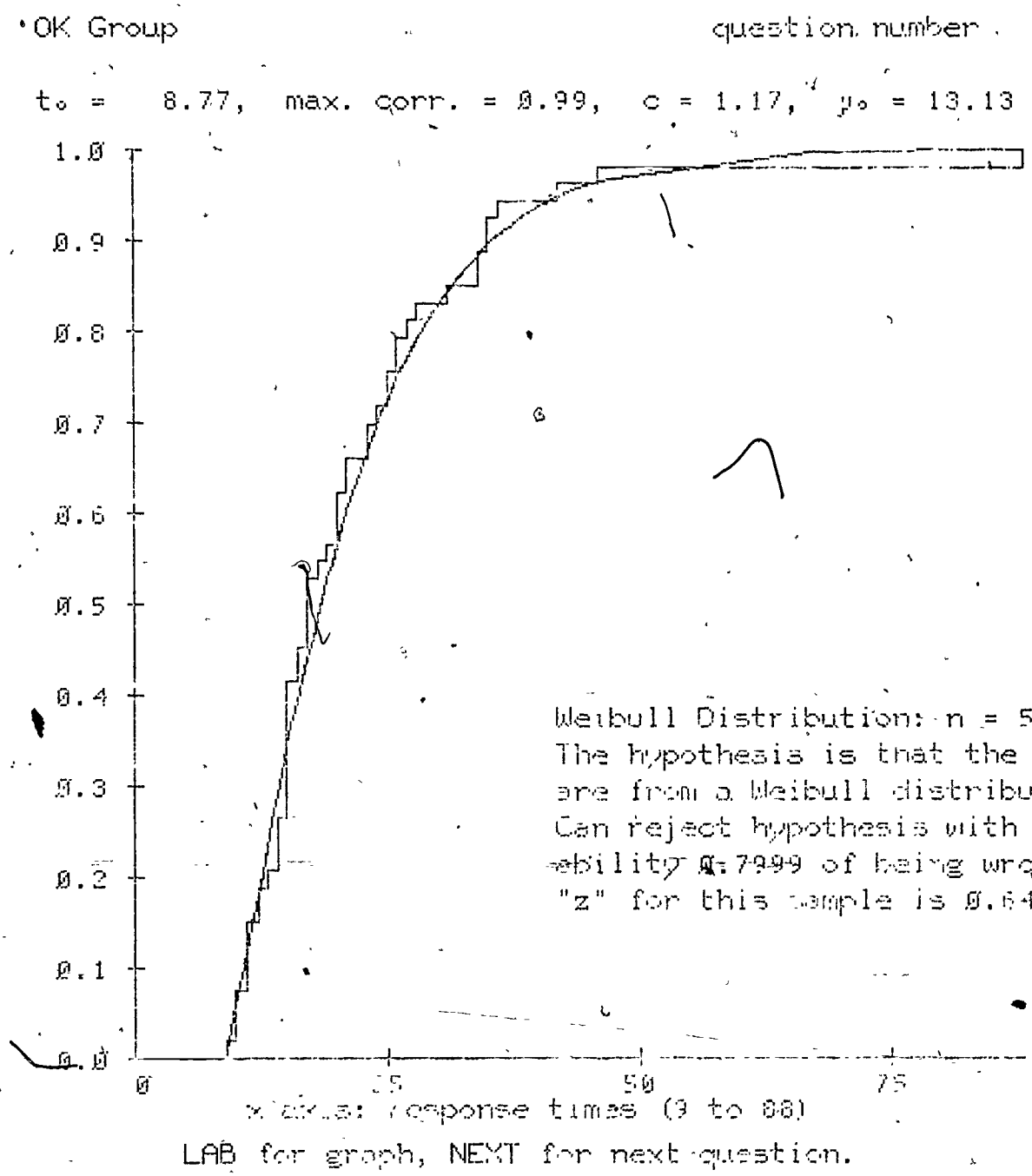


Figure 19 Goodness of fit test for the time-score data and Weibull distribution function

random number generator to supply as matrix elements. It may be impossible to determine a unique t_o-value for these items. The Weibull distribution requires a location parameter t_o to be estimated from the observed data, and it is impossible to estimate such a value when a single t_o really does not exist. Maybe this is the reason why the time-score data from these items don't fit Weibull distributions so well, in comparison with gamma distributions that does not require a unique location parameter t_o.

In the posttest, the different degrees of difficulty caused by a choice of different sets of numbers became negligible. Students had already learned the simple matrix operations and had plenty of opportunities to practice them before taking their posttest. Therefore the discrepancy among t_os, varying from item to item due to the difficulty of calculation would have been minimized and became negligible also. That is probably why Kolmogorov-Smirnov p-values for the Weibull distribution of the posttest improved a great deal as shown in Table 19.

7.2 Exercises in Matrix Algebra Test that Require Only Mechanical Practice

The matrix multiplication lesson includes eight sections with exercises at the end of each instruction. These sections are as follows.

1. Multiplication of A and B
2. $AB \neq BA$
3. Scalar product
4. Matrix product
5. Quadratic form
6. The principles of matrix operation
7. Diagonal matrix
8. Scalar matrix and Identity matrix

Each exercise has the following format where all elements in matrices are supplied by the random number generator.

All items in each exercise are very easy, straightforward examples of what they have learned in the previous instruction. Therefore each problem involves only mechanical calculation rather than requiring heavy reasoning or thinking. As in Figure 20, each exercise requires simple repetition of calculating a scalar product, and hence a strong similarity can be seen to Rasch's model in which the distribution of time taken to read a passage of N words follows the two-parameter gamma distribution. The repetition of N mechanical calculations corresponds to reading N words, we think.

The time data for a student's first try only were sorted out and goodness of fit testings were processed. A summary is given in

$$A = \begin{bmatrix} -8 & 0 & 0 \\ 0 & 9 & 0 \\ 0 & 0 & 3 \end{bmatrix} \quad B = \begin{bmatrix} 7 & 8 & 4 \\ 1 & 10 & 1 \\ 3 & 9 & 10 \end{bmatrix} \quad C = \begin{bmatrix} -5 & 7 \\ 9 & -1 \\ -6 & 10 \end{bmatrix}$$

$$D = \begin{bmatrix} -1 & 6 & 7 \\ -9 & 9 & -4 \end{bmatrix} \quad E = \begin{bmatrix} -6 & 3 & -1 & 2 \\ 10 & -5 & 8 & -5 \end{bmatrix} \quad F = \begin{bmatrix} 3 & 5 \\ 2 & 7 \end{bmatrix}$$

$$w = \begin{bmatrix} -7 \\ -7 \\ 3 \end{bmatrix} \qquad u' = \begin{bmatrix} 4 & -6 & -6 \end{bmatrix}$$

Answer the following questions:

Choose a number to select a problem;

1. w'u
2. u'w
3. w'w
4. u'u

Figure 20 An example of the excercises in the matrix multiplication lesson

Table 20. The time-score data from these exercise sections fit the gamma distribution better than the Weibull distribution.

Table 20

p-values for Weibull and Gamma: Exercises

Section of Exercise	p_W	p_G	N
e02.1	.54	.82	74
e02.2	.25	.47	61
e02.3	.45	.57	67
e02.4	.68	.95	53
e02.5	.98	.98	16
e02.6	.88	.90	39
e02.7*	---	---	--

Note:--Average p-values: .63 for p_W and .78 for p_G.

*Data in this section was lost.

Recall that the items generated by the random number generator, 2, 4, 6 and 9, in the matrix algebra test had a tendency for their time data to show favor to the gamma distribution in the pretest. But in this case, students took exercises after completion of the related instruction, so the argument about the difficulty of determining the minimum required time to respond to a given item in the pretest situation cannot be applied to the situation here. Weibull became a better fit for Items 2, 4, 6 and 9 in the posttest situation. We will need another reasoning to explain why gamma is better than Weibull in the exercises after the instruction.

The two-parameter gamma distribution is a convolution of k independent variables which each follow identical negative exponential distribution functions, and the negative exponential distribution can be obtained by considering the waiting time between arrivals in a random process. Rasch (1961) constructed his oral reading model (word reading model) by drawing an analogy with a simple problem in telephony: the occurrence of a telephone call as a random event, determined by a "calls intensity" parameter which is stable over a certain length of time. In the exercise unit shown in Figure 20, each question involves four simple calculations, and three out of them require multiplying the ith element of one vector by ith element of the other vector and the last calculation involves adding up the three

results of multiplication to get the scalar product. We view these operations as being simple and mechanical enough to identify them with reading the k words which were used in Rasch's word reading model.

Exercises in a Problem Solving Style. Three problem solving style exercises were implemented in the lesson teaching eigenvalues and eigenvector problems. For example, one problem is aimed at guiding a student, step by step, to the goal of calculating eigenvectors of a 2 x 2 matrix. There are four or five stages required to arrive at the final answer and all stages are linearly related, so that previously given stages are required as prerequisites to understanding a later stage. Therefore this type of exercise violates the assumption used in deriving gamma distributions. Note that Weibull distributions don't require such a restrictive assumption and hence have wide applicability and flexibility to more general examples according to Weibull (1951). We predict that the time-score data from exercises in a problem-solving style will fit Weibull distributions very well, and Tables 21 and 22 back up our prediction.

Table 21

p-values for Weibull and Gamma:
Problem-solving Type Exercises

Unit Names	p_W	p_G	N
e05.1	.98	.86	31
e05.2	.93	.05	30
e05.3	.99	.84	29

Table 22

Weibull Parameters for Problem-solving Type Exercises

Unit Names	t_o	c	μ_o	Average time*
e05.1	1.673	.880	11.36	18.39
e05.2	3.758	.790	9.96	24.77
e05.3	3.593	.962	12.14	16.72

*Unit of time is 10 seconds.

The c's for these three exercises are smaller than 1. Since, despite this fact, the average times are very short, it may be that the abundance of hints given during exercises allows many students to speed up toward reaching their given goal.

7.3 Instructional Units or Areas in Matrix Algebra Lessons

Matrix algebra lessons were divided into nineteen small segments or instructional units and the elapsed time to complete each instructional unit was collected. Since these lessons did not adopt a mastery learning strategy, it was impossible to collect mastery time which is the time needed to master a given instructional unit, so the first completion time of each unit was used for analysis. The results of Kolmogorov-Smirnov testing are summarized in Table 23.

Table 23

p-values from Kolmogorov-Smirnov Tests: Matrix Areas

Areas	Content	N	p_W	p_G	Average Time*
i01.1	Simple operations	128	.09	.00	10.5
i01.2	Use of system calculator	134	.30	.01	2.0
i02.1	Multiplication of matrices A, B	135	.30	.50	6.1
i02.2	AB ≠ BA	123	.73	.53	1.8
i02.3	Scalar product	114	.02	.01	1.0
i02.4	Matrix product	116	.14	.13	1.3
i02.5	Quadratic form	122	.33	.56	3.1
i02.6	Properties of operations	109	.21	.12	1.7
i02.7	Diagonal matrix	104	.66	.29	2.3
i03.1	Identity matrix	105	.14	.19	4.9
i03.2	Determinant	103	.50	.48	13.5
i03.3	Evaluation of determinant	101	.64	.32	7.1
i03.4	Cofactors	100	.81	.68	8.9
i03.5	Properties of determinant	98	.62	.72	9.9
i03.6	Adjoint and inverse matrix	102	.95	.76	11.9
i04.1	Rotation of axes	73	.00	.02	.8
i04.2	Orthogonal transformation	52	.56	.79	19.4
i04.3	SSCP matrix	48	.82	.99	19.1
i05.1	Eigenvalues and eigenvectors	72	.71	.83	14.8

*Unit of time is rounded to the nearest minute.

The number of areas whose p-value is larger than .20 for Weibull is 14 and that for gamma is 12, but the number of areas whose p-value is larger than .40 is 10 in each case. The average p-value, $\overline{p}_W$ and $\overline{p}_G$ are .448 and .417 respectively. It is difficult to say which distribution is more suitable to our data, because five areas are classified in Category 1, while seven areas are in Category 4. The combined Category 1, 2 and Category 3, 4 respectively, include eight areas each.

Three instructional units, i01.1, i02.3 and i04.1 don't fit either of the distributions while others fit both Weibull and gamma pretty well. A close examination of all area data from the matrix lesson revealed that 20 to 30 percent of time data from some areas were not the right kind of data that we were interested in. Before the revision of the lessons was made, quite a number of students complained about the lack of flexibility in the original version of the lessons. The original lesson did not have an index page, so if a student starts a lesson, then he/she was forced to go through the lesson without changing the topic until the end. Therefore students were more concentrated on studying and many stayed on the same lesson until they finished it all. In the new version, some students got out of one section before they finished it, and they went back to the index page at the middle of instruction by pressing the key that is always available at any page. The time data used in Table 22 were not exactly the completion time of each section since 20 to 30 percent of the students did not complete some areas.

The time data of the old version fit Weibull very well. The first three lessons (definitions and simple operations, matrix multiplication and determinant, cofactors and inverse) were divided into nine instructional units (areas) and the time data from these nine areas were analyzed. Their fit to the Weibull distributions was very good, with the average p-value being .80. If the data is fairly clean, then a small segment of instructional unit fits the Weibull distribution very well.

It was interesting to note that one area which was given twice during the course showed a remarkably low p-value for the second presentation. When students studied this area for the first time, the p-value was .95, but on the second time, it was only .03.

7.4 The Lessons of Special and General Vehicle Training Program at Chanute Air Force Base

The Chanute AFB CBE project developed 34 lessons to teach repairing and maintenance of various vehicles on the PLATO system. They also developed their own computer managed instruction system and

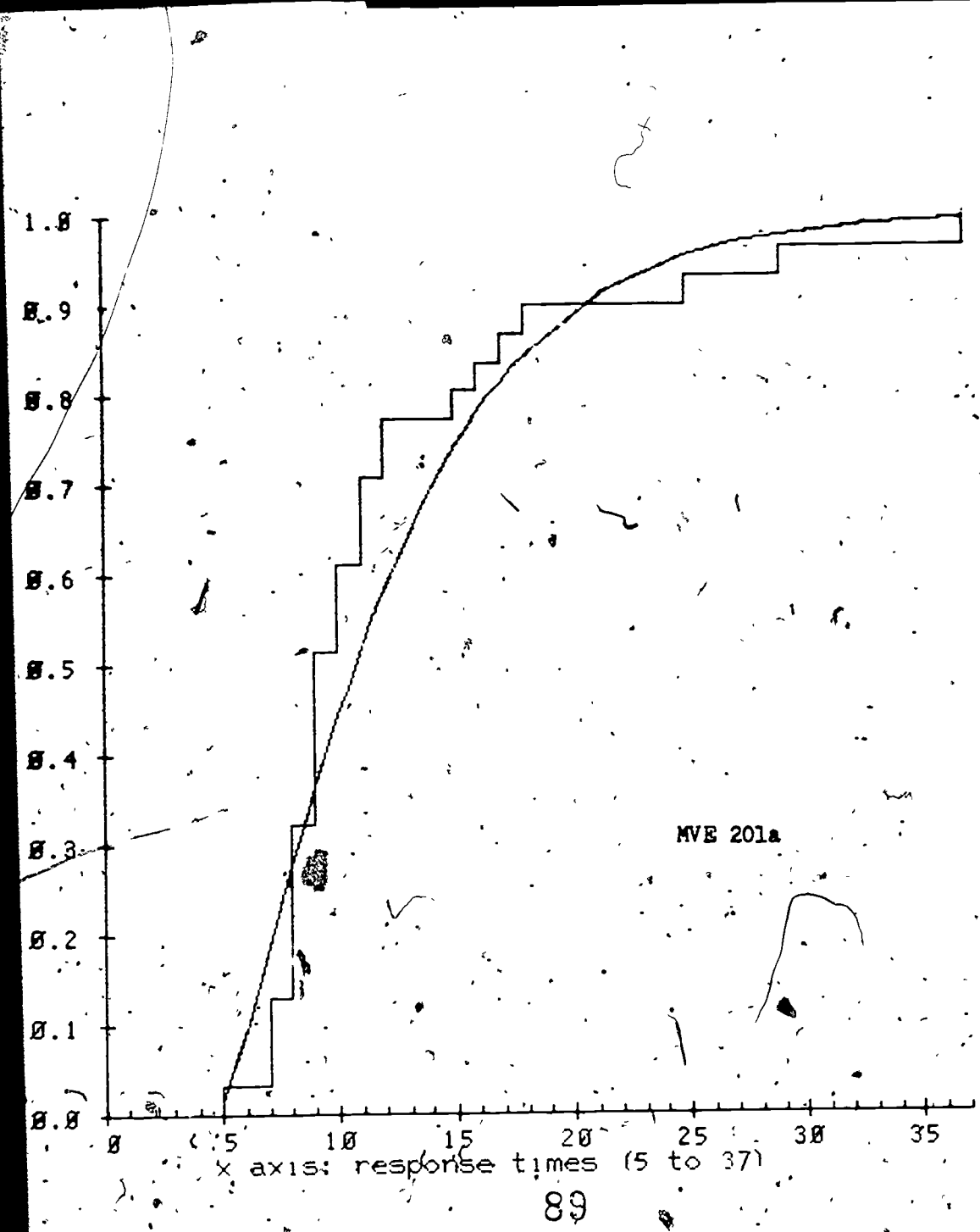
student router. Their unit of measured time was rounded to the nearest minute. Some tests required a few minutes to complete for all students, while other tests had an average time of longer than 10 minutes. About 90 percent of the examinees needed about three minutes to complete many Mastery Validation Exams, and hence rounding to the nearest minute was too rough to analyze these time data, so we had to throw them away.

The time score from the lessons were much better than those from the Master Validation Exams, but a few lessons had a very short average time needed to complete and master the lessons. For example, Lesson mve 201a requires an average time of only 12.6 minutes to master the lesson, yet 72 examinees studied the lesson. If the unit of time were the nearest second, then the p-values from Kolmogorov-Smirnov would have become larger. The plotting of a stepfunction (observed time data) together with a smoothcurve (Weibull distribution function) in Figure 21 has a very large increase of height around the mean value of 12.55 minutes. That affects the z-value. These two plottings look like a fairly close match intuitively, while the average time needed to master Lesson mve 202a is 189.63 minutes and the steps of the observed curve are very fine in Figure 22. The correlations of average times and p-values from Weibull-fitting over 27 Chanute lessons is .57, p-values from gamma-fitting over 27 lessons is .34. Therefore it will be wise to take a finer unit of time in educational research utilizing time scores. Since Chanute lessons used a mastery learning strategy, two kinds of time data were available; one is the first completion time for a given lesson and the second is the time needed until a student achieves a given criterion of mastery at the end of the lesson test, Master Validation Exam.

The following tables, 25 and 26, present a summary of Kolmogorov-Smirnov tests for gamma and Weibull distributions. Appendix C explains the content area that all Chanute lessons were aimed at, and average time for each lesson. Table 24 shows the Weibull parameters.

The average p-value for Weibull is .46, and that for gamma is .51. These values are almost the same as the average p-values for the areas in matrix algebra lessons, but they are not so high in comparison with those of test items and exercises in matrix algebra (refer to Tables 14, 18, 19 and 21).

Although both of the average p-values are only around .50, about 80 percent of the lessons have p-values of larger than or equal to .20 in mastery time (time needed to achieve a given mastery level), which is a satisfactory result. Table 27 shows that gamma is slightly better than Weibull. Since the gamma distribution that has been considered here is a two-parameter distribution without a location parameter, it is bothering to note that gamma is better for this case. Because learning the material written in a whole lesson is not a matter of simple process, or an event that can be explained in



21a "Goodness of fit" for WEibull distribution of lesson MVE201a

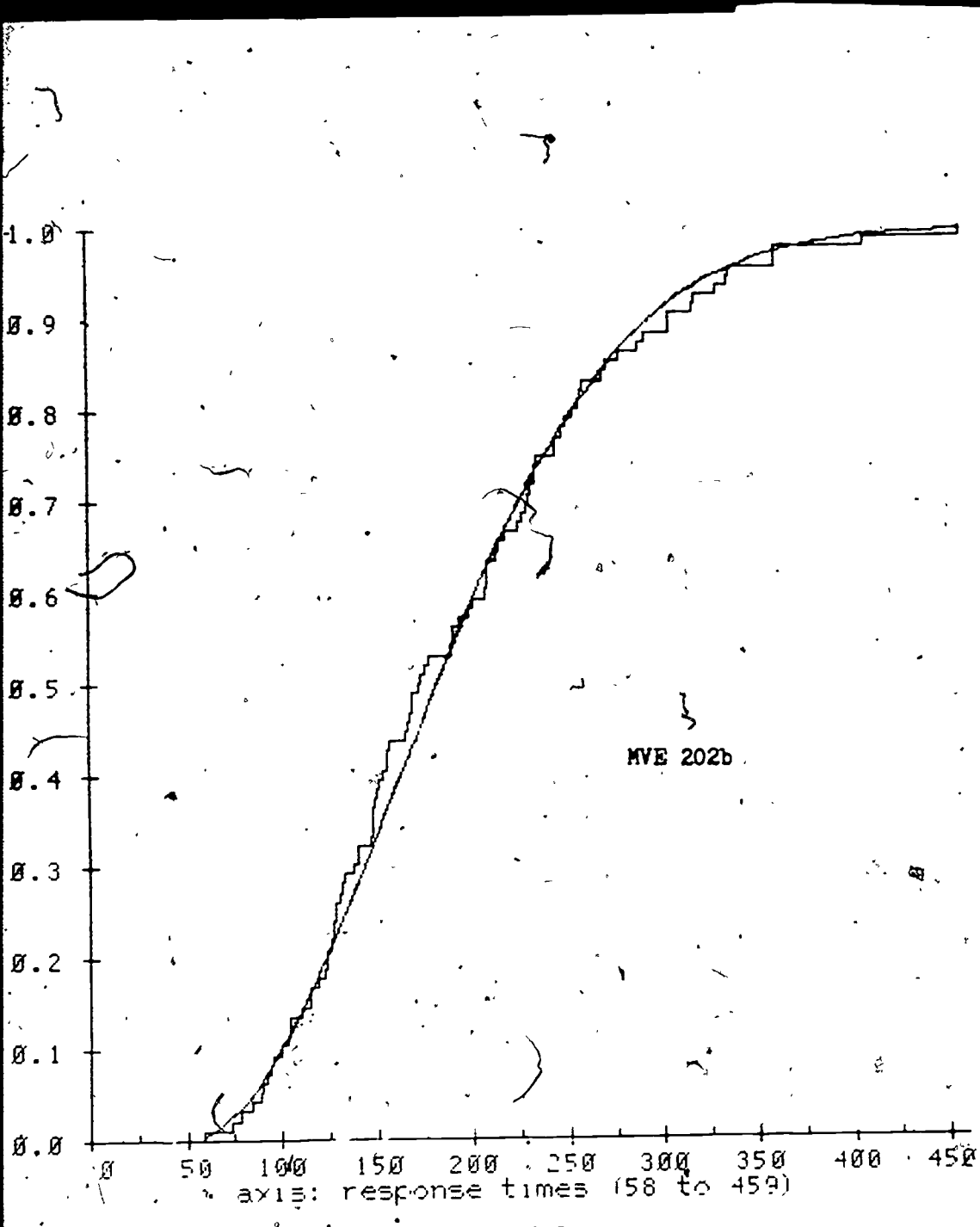


Figure 21b "Goodness of fit" for Weibull distribution of lesson MVE202b

Table 24

Three Weibull Parameters and the Maximum Correlation for Mastery Time

	t	m.c	c	
1)	6.87	.98	1.61	30.03
2)	15.18	.99	1.55	21.54
3)	16.21	.99	1.87	32.65
4)	4.72	.97	1.25	8.06
5)	.00	.98	1.53	48.32
6)	48.67	.99	1.95	160.86
7)	4.49	.99	2.53	11.32
8)	8.63	.99	1.81	105.64
9)	35.38	.98	1.79	126.09
10)	5.38	.98	1.67	42.22
11)	14.85	.99	1.77	42.19
12)	3.13	.99	1.66	20.72
13)	4.19	.99	1.81	46.82
14)	13.21	.99	1.29	27.16
15)	.00	.99	1.81	36.65
16)	4.27	.99	1.85	11.24
17)	.57	.99	2.10	14.37
18)	6.38	.99	1.52	18.40
19)	7.93	.99	2.11	74.52
20)	11.50	.99	1.75	44.13
21)	6.32	.99	2.85	20.37
22)	20.55	.99	1.63	79.59
23)	.76	.99	1.21	15.91
24)	.00	.99	1.75	14.27
25)	4.33	.99	1.76	31.10
26)	12.22	.99	1.53	45.01
27)	2.53	.97	1.55	12.96

Table 25

Kolmogorov-Smirnov Tests for Chanute Data

Lesson	Mastery Time			Completion Time		
	c	z	N	c	z	N
1)	0.1248	1.1777	85	0.2716	0.9983	85
2)	0.6094	0.7606	83	0.7286	0.6895	83
3)	0.3349	0.9440	85	0.4193	0.8811	85
4)	0.2123	1.0587	72	0.0156	1.5574	74
5)	0.6653	0.7274	86	0.8500	0.6106	86
6)	0.7731	0.6621	96	0.5205	0.8146	95
7)	0.0624	1.3168	81	0.0712	1.2914	81
8)	0.6711	0.7240	86	0.7874	0.6530	85
9)	0.4919	0.8328	89	0.4251	0.8771	89
10)	0.5622	0.7890	78	0.7288	0.6894	77
11)	0.8667	0.5981	75	0.4350	0.8703	75
12)	0.2714	0.9987	80	0.2121	1.0590	80
13)	0.6780	0.7198	67	0.3768	0.9117	67
14)	0.8316	0.6236	87	0.6781	0.7198	87
15)	0.7671	0.6658	77	0.6552	0.7334	77
16)	0.3222	0.9543	71	0.2360	1.0333	71
17)	0.1387	1.1550	67	0.2167	1.0538	67
18)	0.1438	1.1471	72	0.1344	1.1618	72
19)	0.9095	0.5627	62	0.9189	0.5538	62
20)	0.1212	1.1838	76	0.1682	1.1124	75
21)	0.4265	0.8761	59	0.4176	0.8823	59
22)	0.3812	0.9084	93	0.1583	1.1260	93
23)	0.0988	1.2264	73	0.0743	1.2832	71
24)	0.4985	0.8286	67	0.6039	0.7638	67
25)	0.4089	0.8884	70	0.8339	0.6220	70
26)	0.9599	0.5062	70	0.9929	0.4289	70
27)	0.0938	1.2370	69	0.8860	0.5830	68

'goodness of fit' testing for Weibull

Table 26

Kolmogorov-Smirnov Tests for Chanute Data:

"Goodness for fit" for Gamma

Lesson	Mastery Time		N	Completion Time	
	p	z		p	z
1) mve103	0.0643	1.3110	85	0.1260	1.1757
2) mve104a	0.4537	0.8577	83	0.5184	0.8159
mve104b	0.8373	0.6196	6	1.0000	0/0
3) mve105	0.5619	0.7891	85	0.7193	0.6952
4) mve201a	0.0361	1.4169	72	0.0544	1.3425
5) mve201b	0.6318	0.7472	86	0.5357	0.8052
6) mve202a	0.8646	0.5998	96	0.9889	0.4449
7) mve202b	0.1554	1.1300	81	0.1681	1.1130
8) mve204	0.9260	0.5468	86	0.5726	0.7827
9) mve205a	0.5686	0.7851	89	0.7117	0.6997
10) mve205b	0.1923	1.0818	78	0.9316	0.5410
11) mve206a	0.7981	0.6460	75	0.5745	0.7815
12) mve206b	0.6267	0.7503	80	0.6043	0.7636
13) mve206c	0.7859	0.6540	67	0.5758	0.7307
14) mve207	0.2470	1.0222	87	0.3833	0.9069
15) mve301	0.8557	0.6064	77	0.8879	0.5814
16) mve303	0.5185	0.8159	71	0.3720	0.9153
17) mve304	0.6996	0.7070	67	0.7225	0.6932
18) mve305	0.3119	0.9629	72	0.2130	1.0580
19) mve307	0.6292	0.7488	62	0.8028	0.6430
20) mve308	0.2856	0.9858	76	0.4262	0.8764
21) mve401	0.8128	0.6364	59	0.0232	1.4924
22) mve402	0.2362	1.0331	93	0.5650	0.7873
23) mve403	0.2761	0.9944	73	0.1003	1.2232
24) mve404	0.8396	0.6180	67	0.9764	0.4778
25) mve405a	0.3883	0.9032	70	0.9962	0.4087
26) mve405b	0.6758	0.7211	70	0.9895	0.4430
27) mve405c	0.0561	1.3366	69	0.2597	1.0097

parallel to a Poisson process, or like Rasch's words reading model, it is probable that we will have to investigate the composite distribution model for Weibull distribution, instead of a single distribution. An r-component composite Weibull distribution is defined as $F_x(x) = F_j(x)$, $S_j \leq t \leq S_{j+1}$ for $j = 0, 1, 2, \ldots r$. Further mathematical discussion will be found in Mann et al. (1975). In future work, we will have to analyze carefully a whole task of instruction in a lesson and divide it into finer tasks. The time-score data from each task unit (or segment of instruction, or area) can be represented by a Weibull distribution. If a lesson is of k tasks, then a k-composite Weibull distribution will be the distribution representing the whole lesson. Since it is impossible to investigate further along this line with Chanute lessons, we will work with matrix area data (after cleaning up the messy data) in the near future.

Table 27

27 Vehicle Maintenance Training Lessons, p-values from Kolmogorov-Smirnov Testing for Weibull and Gamma Distributions

		p > .20	p > .40	p > .50
Weibull	1*	20 (74%)	15 (56%)	12 (44%)
	2**	21 (78%)	16 (69%)	12 (44%)
Gamma	1*	22 (81%)	18 (67%)	18 (67%)
	2**	23 (85%)	17 (63%)	16 (59%)

* Time needed to complete a lesson.

** Time needed to reach a given mastery level.

Although the mastery time obtained from Chanute lessons did not fit Weibull distributions quite as well as time-score data from matrix test items did, the shape parameter in this context c has an important relationship with one of the current topics in educational measurement: the problem of false negatives and false positives of criterion-referenced tests. The detailed analyses and discussion of the role of the shape parameter c will be given in the next section. The table of Weibull parameters for mastery time data from Chanute lessons was given in Table 24.

Revised Chanute Lessons. After the initial data (the result of the previous section was based on this data) from all lessons in the vehicle maintenance training course were collected and analyzed, seven lessons were selected for further modification and revisions. A year later, the first completion time of these polished, revised lessons were collected and tested for goodness of fit with Weibull distributions. The changes that were made were quite extensive and average times of the lessons became quite different from the original version of seven lessons; some got longer but others got shorter. But the p-values from Kolmogorov-Smirnov tests became much larger than the original ones. These values are shown in Table 28.

Table 28

Comparison of p-values from Kolmogorov-Smirnov Tests for the Original Lessons and their Improved Versions

Lessons	Original p-value	Revised p-value
202b	.07	.90
204	.79	.54
207	.67	.68
301	.66	.73
307	.91	.91
308	.17	.79
401	.42	.79

Table 28 might suggest that the time data from the more polished, improved lessons fit Weibull distributions better than those from the less polished, original version of lessons. The less polished lessons usually contain ambiguous explanations, typographical errors, inappropriate feedbacks or improper amounts and quality of help. Eliminating such distractions that affected a student's pace of learning badly, especially for those who were not so bright, or for those who knew nothing about the material, might have caused better fit with Weibull distributions. This fact implies that the study of CRR will lead us to identify the quality of feedbacks, appropriateness of the help branch in terms of using qualitative analysis methods. We believe that our research will be very useful to the area of instructional design in a practical sense; we can provide a quantitative tool to instructional designers who are mostly artists.

8. THE CORRELATES OF PROBABILITIES OF MISCLASSIFICATION BY CRITERION-REFERENCED TESTS

In this section we explore what variables are associated with erroneous decisions--calling a non-master a master (false positives, F+) and calling a master a non-master (false negatives, F-) based on the criterion-referenced tests of the Chanute AFB CBE Project. The Weibull shape parameter c turned out to be a prominent predictor of the estimated probabilities p(F+) and p(F-).

Another thrust of this section is the definition of a new index, dubbed the "efficiency index," which we believe to be a reasonable measure of the quality of a lesson. A factor analysis using 18 variables (including p(F+), p(F-), p(F+ or F-), α_{21}, failure rate, the three Weibull parameters, the distance between the optimum cutoff point and the mean, etc.) along with this efficiency index yielded a distinct factor loading only this variable and c.

8.1 Beta Binomial Model

Criterion-referenced testing (CRT) has gained much attention from educational measurement and testing specialists in recent years. The object of criterion-referenced testing is not to distinguish finely among subjects, but to classify subjects into mastery and non-mastery groups. Hence the accuracy of judging non-mastery or mastery status of examinees becomes the main concern.

Since criterion-referenced tests are commonly used in situations where students are expected to achieve the level of mastery, say 90 percent correct, the observed scores become a bounded variable. If there are subjects with true scores near the "ceiling" or the "floor," it becomes implausible to assume that the errors of measurement are distributed independently of true scores for those near the boundary.

Lord and Novick (1968) argue about the plausible distributional forms of observed CRT scores and true scores in Chapter 23 of their book, "Statistical Theories of Mental Test Scores." We will follow their steps and adopt the binomial error model for CRT scores. The binomial error model assumes that if each MVE test is aimed at measuring the learning level of a topic taught in the Vehicle Training Course of the Chanute AFB DBE Project, for instance, then all items in the test must measure the same task. In other words all items in a test have one and only one common factor with 0-1 scoring. Suppose there is a pool of items measuring the same task, and taking an item out of the pool is an independent event, that is, answering the earlier items on the test does not affect the ability of a student to answer

later items correctly, then we can formulate the distribution of raw scores x by a binomial distribution with parameter θ in which θ is the proportion of items that a student would answer correctly over the entire pool of items. If T is a fixed true score and e is an error of measurement, then the raw score x can be expressed as the sum of the two, $x = T + e$, and θ is given by

$$\theta = T/n$$

where n is the number of items in the test. Let $g(x|\theta)$ be the binomial distribution of x at any given true ability level θ, then the conditional distribution $g(x|\theta)$ can be given by

$$g(x|\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x} \qquad x = 0,1,\ldots,n.$$

It is interesting to note that this model does not pay attention to item differences. The traditional measurement indices such as item difficulty or items discriminating index are not the major concern in the binomial error model. Rather, finding out how accurately a test can estimate an examinee's pass or fail status with respect to a given mastery criterion is the main concern of the model.

Keats and Lord (1962) investigated the relationship between the distribution of test scores, observed and true scores. The test scores could be adequately represented by the hypergeometric distribution f(x) with a negative parameter and the true score distribution could be represented by the two-parameter beta distribution g(0).

$$g(\theta) = \theta^{a-1}(1-\theta)^{b-n}/B(a,b-n+1)$$

where $a > 0$ and $b > n-1$. And also

$$g(x) = \int_0^1 \frac{\theta^{a-1}(1-\theta)^{b-n}}{B(a,b-n+1)} \cdot \binom{n}{x}\theta^{x}(1-\theta)^{n-x}d\theta, \quad x = 0,1,\ldots,n.$$

In classical test theory, the estimation of a true score is given by regressing the true score T on the observed score x, and the equation is given by

$$E(T|x) = \rho x + (1-\rho)\mu_x$$

where ρ is the reliability of the test and μ_x is the mean of test scores.

In the binomial error model, the estimation of a true score is given by a similar equation,

$$E(T|x) = \alpha_{21}x + (1 - \alpha_{21})\mu_x, \qquad x = 0,1,\ldots,n.$$

where α_{21} is the ratio of number-correct true-score variance to observed-score variance and is given by

$$\frac{\alpha_T^2}{\alpha_x^2} = \frac{n}{n-1}\left\{1 - \frac{\mu_x(n-\mu_x)}{n\,\alpha_x^2}\right\} = \alpha_{21}$$

Table 29 is the summary of information from the Mastery Validation Exams at Chanute.

The mastery level of Master Validation Exams (MVE) of the 34 lessons in the Chanute AFB CBE Project was set at a level of 80 percent; although it is hard to prove that 80 percent is the most appropriate level for their program. Block (1972) showed in his experimental study that attainment of a 95 percent mastery level maximized student learning of cognitive tasks in his matrix algebra course, while an 85 percent level maximized learning as characterized by affective criteria.

Since Chanute's 34 lessons are designed to be "homogeneous" with respect to content and teaching style, all lessons are written under the same principle with the same tutorial logic, although the subject matter in each lesson is different. Therefore Chanute's lessons are not linearly related and the content difficulty of the lessons is not hierarchically ordered as it would be in teaching mathematics, arithmetic, or foreign languages. If the lessons are linearly related, setting a mastery level for the earlier instructional units should be higher than those of the later instructional units. If the goal of the second unit is the attainment of an 85 percent mastery level, then the mastery level of the first unit might be 90 percent, or some other level higher than 85 percent. Since there is no analytical technique to provide the optimal level of mastery learning, definite statements about the determination of ideal mastery levels cannot be made at this time. Linn (1978) provides an excellent discussion about the topic of "setting standards."

Table 29

The Summary of Simple Statistics of Mastery Validation Exams

test	mean	SD	items	α21	N
mvel03	7.388	1.124	8	0.6321	85
mvel04a	11.892	0.442	12	0.4910	83
mvel04b	10.120	1.728	11	0.8018	83
mvel05	7.706	0.737	8	0.5470	85
mve201a	9.474	0.973	10	0.5254	76
mve201b	8.907	1.325	10	0.4951	86
mve202a	16.186	2.934	20	0.6753	97
mve202b	9.720	0.634	10	0.3573	82
mve204	8.557	1.681	10	0.6253	88
mve205a	6.767	1.558	9	0.3470	90
mve205b	8.110	1.736	10	0.5457	82
mve206a	12.038	1.574	13	0.6942	78
mve206b	15.250	1.619	17	0.4259	80
mve206c	19.257	1.151	20	0.4841	70
mve207	3.761	1.124	5	0.3287	88
mve301	8.727	1.501	10	0.5635	77
mve303	17.380	2.257	20	0.5824	71
mve304	9.209	1.366	10	0.6771	67
mve305	7.458	0.934	8	0.4806	72
mve307	14.683	1.522	16	0.5101	63
mve308	9.037	1.170	10	0.4045	82
mve401	9.254	1.015	10	0.3673	63
mve402	14.138	2.335	17	0.5988	94
mve403	8.095	2.487	10	0.8340	84
mve404	4.254	0.876	5	0.2166	67
mve405a	9.169	1.069	10	0.3701	71
mve405b	8.329	1.991	10	0.7208	70
mve405c	9.087	1.222	10	0.4934	69

Mastery levels are usually set by instructors or the author of a lesson, but the decision of mastery and non-mastery is based on examinees' observed test scores. The score that is used to decide mastery and non-mastery is called the "cutoff." Mastery and non-mastery statuses ought to be defined on the basis of true ability θ, not observed test scores x that are subject to measurement errors. If true ability were known, there would be no incorrect classifications. Unfortunately, true scores are impossible to obtain in practice, so we have to find a way to minimize misclassification.

There are four kinds of classifications: (1) an examinee's true ability θ is higher than a given mastery level θ_o and the observed score x is higher than the cutoff score c, that is $A = \{\theta \geq \theta_o$ and $x \geq c\}$; (2) θ is lower than θ_o and x is also lower than c, that is $B = \{\theta < \theta_o$ and $x < c\}$; (3) θ is lower than θ_o, but x is larger than c, $F_+ = \{\theta < \theta_o$ and $x \geq c\}$; (4) θ is higher than θ_o, but x is lower than c, $F- = \{\theta \geq \theta_o$ and $x < c\}$. Figure 22 shows these four conditions.

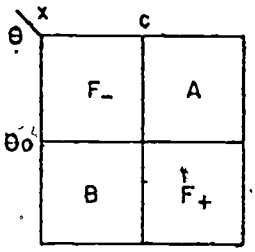


Figure 22 Classification Table

θ : true ability, x: observed score

θ_o: true mastery level

c : observed cutoff

Probability of these events will be denoted by P(A), P(B), P(F+) and P(F-) respectively

Millman (1975) and then Novick and Lewis (1975) reported the percentage of students expected to be misclassified for a given cutoff with various numbers of test items. Millman used the binomial error model, but Novick and Lewis used the Bayesian beta binomial error model.

According to Millman's calculations, the percentage of students expected to be misclassified at 80 percent mastery level using a 10-item test could be as high as 53 percent.

Emerick (1972) and Huynh (1976) considered the loss ratio Z of F- to F+ as a means of controlling misclassification, especially false advancement. If later instructional units require the knowledge and skill acquired in earlier units, false advancement will be a problem. The loss ratio of 10 implies the event F- is ten times as serious as the event F+. Since F- stands for the event in which a student has really mastered the given instructional unit but his/her observed score happens to be lower than the cutoff, retaining such a student in the same unit is not efficient. If the instructional units are fairly independent from one to another, as are lessons in the Vehicle Training Program at Chanute Air Force Base, then an appropriate loss ratio would be 1, or at least it is not necessary to set it as high as 10.

Huynh (1976) proposed an evaluation of the cutoff score that minimizes the occurrence of misclassifications for a given loss ratio. With his cutoff score, the loss ratio of having a false positive to having a false negative stays the same, say 10, while the linear combination of the probabilities of the both events and the loss ratio (the average loss) is minimized. We will discuss in more detail Huynh's method in conjunction with 34 Chanute lessons and their MVE test scores.

8.2 Evaluation of the Optimal Cutoff Scores

Huynh derived the optimal cutoff c_o of a test for a given mastery level θ_o and loss ratio Q so as to minimize the average loss function R(c) which is the following linear combination of the probabilities of false positive and false negative:

$$R(c) = P(F+) + Q\ P(F-)$$

It turns out that c_o is the smallest integer such that the incomplete beta function $I_{\theta_o}(a+c_o,\ n+b-c_o)$ is smaller than or equal to $Q/(1+Q)$; where

$$P(c_o) = I_{\theta_o}(a+c_o,\ n+b-c_o) = \int_o^{\theta_o} \frac{\theta^{a+c_o-1}(1-\theta)^{n+b-c_o-1}}{B(a+c_o,\ n+b-c_o)}\ d\theta$$

In order to apply Huynh's result to evaluate c_o, we need the help of a computer to calculate and plot the values of the incomplete beta function for $c_o = 0,1,2,\ldots n$. The PLATO system eases these steps and we can obtain the answer through the program "cutoff." Figure 23 illustrates the procedure to determine the optimal cutoff c_o. The

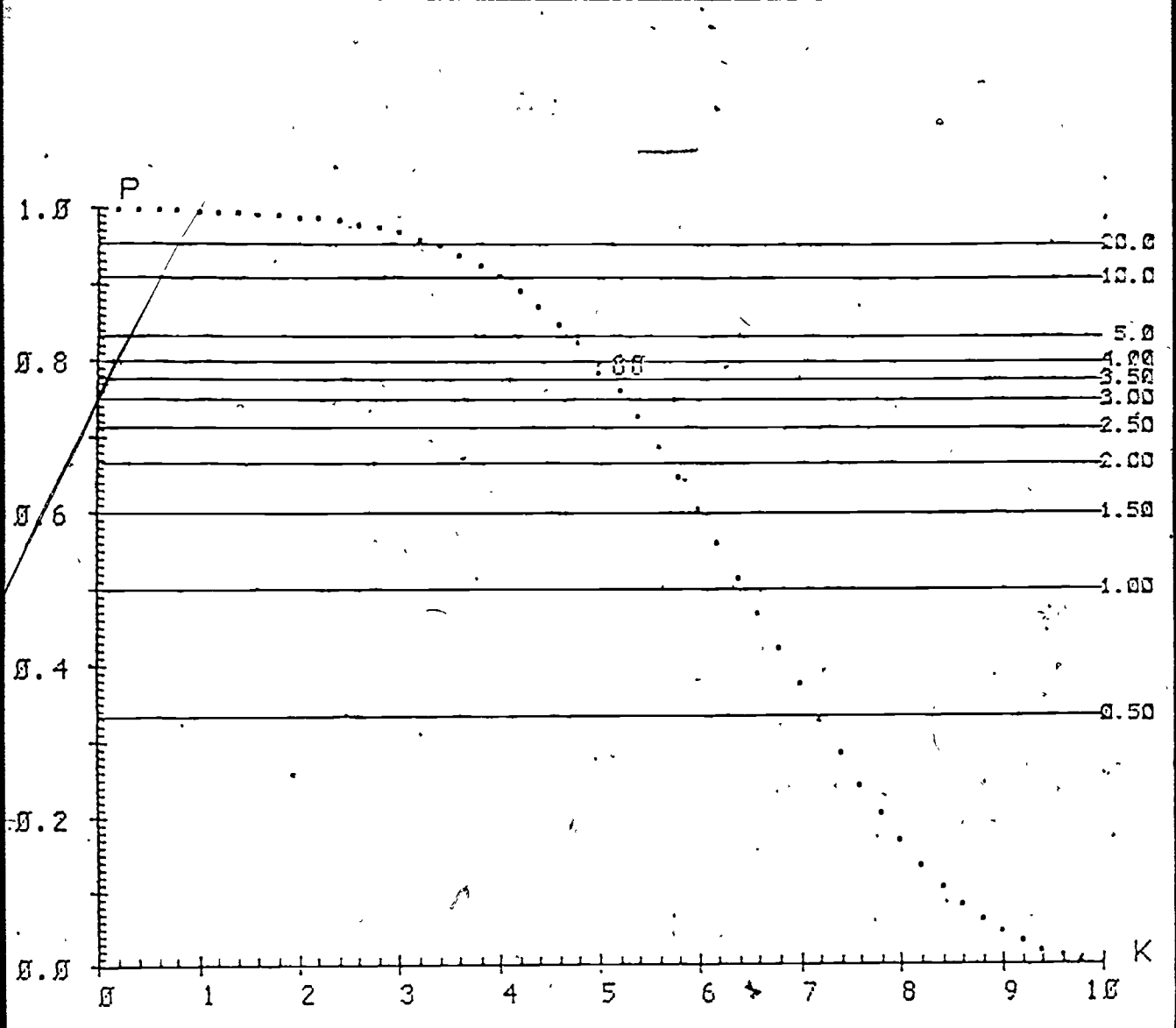


Figure 23 Determining the optimal cutoff C_0 as to minimize misclassification

lesson = MVE201a		subjects = 76		n = 10	
mean = 9.4737		SD = 0.9726		x_{21} = 0.53	
a = 8.5560		b = 0.4753			

parameters a and b are obtained from the mean, standard deviation of the test and the number of items in the test (denoted by n). Table 30 shows the values of incomplete beta function $I_{\theta_o}(i)$, at each point $i = 1,2,\ldots,n$, where a,b are calculated from test scores of mve201a by the formulas

$$a = (-1 + \frac{1}{\alpha_{21}})\mu_x$$

$$b = -a + \frac{n}{\alpha_{21}} - n$$

The curve in Figure 23 is obtained by plotting the points in Table 30. The horizontal lines which are marked by losses 0.5, 1, 2, 3, and 4 in Figure 23 help to evaluate the optimal cutoff which minimizes the average loss R(c) at c_o for the partially known loss ratio Q and a given true mastery level θ_o. Since the contents of all lessons discussed in the Chanute AFB CBE Project deal with independent topics across the lessons and the lessons are not linearly or hierarchically related, a loss ratio of 1 will be reasonable. Note that in Figure 23 the smallest integer value of i for which the curve P(i) goes under the line of loss ratio 1 is 7. Therefore $c_o = 7$ is the ideal cutoff score of the test, mve201a.

Table 30

Ten Points in Figure 23

Item	a+i	n+b-i	I_{θ_o} (a+i, n+b-i)
1	9.556	9.475	0.998
2	10.556	8.475	0.991
3	11.556	7.475	0.969
4	12.556	6.475	0.913
5	13.556	5.475	0.796
6	14.556	4.475	0.608
7	15.556	3.475	0.376
8	16.556	2.475	0.169
9	17.556	1.475	0.045
10	18.556	0.475	0.004

$\theta_o = .80$, Test = mve201a, a = 8.5560, b = 0.4753

It is interesting to note that the cutoff score, c=8, actually used for mve201a in the Chanute training program gives a slightly larger value of the probability of misclassification $R(c) = P(F+) + P(F-)$, where Q=1 than the theoretically derived c_o does, but not for P(F+), probability of false positive, or P(F-), probability of false negative separately.

The probability of event B in Figure 22, $P(B) = P(\theta < \theta_o, x < c)$ can be expressed by a linear combination of beta functions and incomplete beta functions, because

$$P(B) = \int_{\theta<\theta_o} \int_{x<c} P(\theta) f(x|\theta) d\theta dx = \int_{\theta<\theta_o} \int_{x<c} \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)} \binom{n}{x} \theta^x (1-\theta)^{n-x} dx d\theta$$

$$= \frac{1}{B(a,b)} \sum_{i=0}^{c-1} \binom{n}{i} B(a+i,\ b+n-i) I_{\theta_o}(a+i,\ b+n-i).$$

Similarly,

$$P(F-) = P(\theta \geq \theta,\ x<c) = P(x<c) - P(\theta<\theta_o,\ x<c) = P(x<c) - P(B)$$

where

$$P(x<c) = \int_{x<c} \frac{\binom{n}{x} B(a+x,\ n-x+b)}{B(a,b)} dx = \frac{1}{B(a,b)} \sum_{i=0}^{c-1} \binom{n}{i} B(a+i,\ n+b-i).$$

$$P(F+) = P(\theta<\theta_o,\ x \geq c) = P(\theta<\theta_o) - P(B)$$

where

$$P(\theta<\theta_o) = I_{\theta_o}(a,b).$$

Thus, we obtain the following calculation formulas for P(A), P(B), P(F-) and P(F+).

$$P(F+) = I_{\theta_o}(a,b) - \frac{1}{B(a,b)} \sum_{i=0}^{c-1} \binom{n}{i} B(a+i,\ b+n-i) I_{\theta_o}(a+i,\ b+n-i)$$

$$P(F-) = \frac{1}{B(a,b)} \sum_{i=0}^{c-1} \binom{n}{i} B(a+i,\ n+b-i)(1-I_{\theta_o}(a+i,\ b+n-i))$$

$$P(A) = 1-I_{\theta_o}(a,b) + \frac{1}{B(a,b)} \sum_{i=0}^{c-1} \binom{n}{i} B(a+i,\ n+b-i)(I_{\theta_o}(a+i,\ b+n-i)-1)$$

$$P(B) = \frac{1}{B(a,b)} \sum_{i=0}^{c-1} \binom{n}{i} B(a+i,\ b+n-i) I_{\theta_o}(a+i,\ b+n-i)$$

The probability of each misclassification for all available Mastery Validation Exams were calculated and summarized in Table 31.

Since the sum of the probabilities A, B, F+ and F- is 1, the sum of the probabilities of A and B must have the maximum value at c_o where the sum of probabilities F+ and F- reaches the minimum. Since mastery and non-mastery status of examinees are actually determined by the observed cutoff c, the probability, $P(x \geq c)$ is the probability of the observed mastery status. Column 6 in Table 31, headed by P(A or F+), is the estimated probability of passing the mastery criterion judged by the observed scores using cutoff c and cutoff c_o respectively. The success rates in Column 7 are the actually observed percentages of examinees who achieved mastery level, i.e. who obtained scores greater than or equal to c. Also Table 31 indicates that the actually used cutoff scores c produce higher probabilities of misclassification than the theoretically determined cutoff c_os except in a few cases. Since the theoretical cutoffs are determined so as to minimize the average loss R(c), in our case the sum of probabilities of false negative F- and false positive F+, all values in Column 6 of Table 31, P(F+ or F-) are smaller for c_o than for c. The sum of the probabilities of A and F+ is the expected success rate, so this sum matches the observed success rate given in the last column fairly well. If c_o were used as cutoffs for MVE test scores, only 12 lessons would have a probability of observed success less than .90, while 20 lessons have values of P(A or F+) less than .90 when c's are used.

Since the probability of false negative, P(F-) stands for the case that an examinee really mastered the goal of instructional unit but his/her observed score happened to be lower than the used cutoff c, he/she does not really have to repeat the instruction. If efficiency of training in terms of shortening the training time is the main concern, then P(F-) should not be so large. For example, MVE207

Table 31

Estimated Probability of Misclassifications

Test	Cutoff[a]	$P(F_+)$	$P(F_-)$	$P(F_+ \text{ or } F_-)$	$P(A \text{ or } F_+)$	Success rate
mvel03	c_o 6	0.0621	0.0162	0.0783	0.9247	.89
	c^* 7	0.0314	0.0639	0.0953	0.8462	
mvel04a	c_o 7	0.0026	0.0001	0.0026	0.9997	.94
	c^* 10	0.0011	0.0057	0.0068	0.9927	
mvel04b	c_o 9	0.0348	0.0259	0.0606	0.8705	.86
	c^* 9	0.0348	0.0259	0.0606	0.8705	
mvel05	c_o 6	0.0235	0.0094	0.0329	0.9739	.88
	c^* 7	0.0123	0.0399	0.0522	0.9323	
mve201a	c_o 7	0.0357	0.0064	0.0421	0.9788	.90
	c^* 8	0.0238	0.0262	0.0499	0.9472	
mve201b	c_o 7	0.1078	0.0146	0.1223	0.9375	.72
	c^* 8	0.0710	0.0556	0.1266	0.8598	
mve202a	c_o 16	0.1163	0.0624	0.1788	0.6495	.82
	c^* 16	0.1163	0.0624	0.1788	0.6495	
mve202b	c_o 5	0.0055	0.0001	0.0056	0.9998	.98
	c^* 8	0.0031	0.0122	0.0153	0.9853	
mve204	c_o 8	0.0996	0.0503	0.1499	0.7803	.94
	c^* 8	0.0996	0.0503	0.1499	0.7803	
mve205a	c_o 8	0.1428	0.1341	0.2769	0.3612	.79
	c^* 8	0.1428	0.1341	0.2769	0.3612	
mve205b	c_o 8	0.1507	0.0634	0.2141	0.6913	.82
	c^* 8	0.1507	0.0634	0.2141	0.6913	
mve206a	c_o 10	0.0478	0.0184	0.0662	0.9207	.82
	c^* 11	0.0266	0.0535	0.0801	0.8644	
mve206b	c_o 12	0.0606	0.0113	0.0719	0.9708	.82
	c^* 14	0.0305	0.0911	0.1216	0.8608	
mve206c	c_o 13	0.0057	0.0003	0.0061	0.9991	.95
	c^* 16	0.0030	0.0116	0.0146	0.9852	
mve207	c_o 5	0.0965	0.1957	0.2922	0.3070	.91
	c^* 4	0.2878	0.0547	0.3425	0.6393	

Table 31 (cont.)

	Cutoff[a]	$P(F_+)$	$P(F_-)$	$P(F_+ \text{ or } F_-)$	$P(A \text{ or } F_+)$	Success rate
mve301	c_o 8	0.0894	0.0540	0.1434	0.8184	.79
	c 8	0.0894	0.0540	0.1434	0.8184	
mve303	c_o 15	0.1070	0.0266	0.1336	0.8867	.90
	c 16	0.0730	0.0653	0.1383	0.8140	
mve304	c_o 8	0.0471	0.0292	0.0763	0.8922	.82
	c 8	0.0471	0.0292	0.0763	0.8922	
mve305	c_o 5	0.0632	0.0036	0.0668	0.9827	.96
	c 7	0.0247	0.0787	0.1034	0.8691	
mve307	c_o 11	0.0526	0.0056	0.0582	0.9797	.81
	c 12	0.0413	0.0187	0.0600	0.9553	
mve308	c_o 7	0.0732	0.0147	0.0880	0.9601	.63
	c 8	0.0498	0.0578	0.1076	0.8936	
mve401	c_o 7	0.0364	0.0109	0.0473	0.9872	.83
	c 8	0.0252	0.0451	0.0704	0.9328	
mve402	c_o 13	0.1494	0.0395	0.1890	0.7809	.79
	c 14	0.0910	0.0961	0.1871	0.6660	
mve403	c_o 8	0.0771	0.0294	0.1065	0.7048	.79
	c 8	0.0771	0.0294	0.1065	0.7048	
mve404	c_o 3	0.2100	0.0130	0.2230	0.9564	1.00
	c 4	0.1455	0.0840	0.2296	0.8208	
mve405a	c_o 6	0.0560	0.0025	0.0585	0.9919	1.00
	c 8	0.0326	0.0513	0.0839	0.9196	
mve405b	c_o 8	0.0987	0.0419	0.1405	0.7344	.91
	c 8	0.0987	0.0419	0.1405	0.7344	
mve405c	c_o 7	0.0794	0.0123	0.0917	0.9543	.94
	c 8	0.0527	0.0478	0.1005	0.8921	

[a] c_o is the theoretically derived cutoff to minimize $P(F_+) + P(F_-)$. c is the cutoff actually used in the PLATO Service Program at Chanute.

has P(F-) = .1957 so that to 88 x 0.1957 or 17 out of a total of 88 students repeated the same instruction unnecessarily. Of course this is an extreme case and most p values are less than .10 percent, which means that five to eight students repeated the same lesson mistakenly. Table 32 shows the number of students who will be misclassified or were misclassified.

We conclude that most cutoffs of Master Validation Exams used at Chanute were not the best choice. By adopting the theoretically derived cutoff c_o's the probability of misclassifications could have been minimized. Note that P(F+) at cutoff c_o for each MVE except for MVE207 (which has c_o larger than c, while others has the reverse) becomes larger than or equal to the value of P(F+) at cutoff c, while P(F-) showed the reverse phenomenon. The appropriate judgement of which misclassification should be minimized must be made by a test administrator through deciding on the loss ratio Q. We set Q=1 because all lessons were considered to be not related linearly. We have to face the problem of how to put weights on the cases, the increased chance of having students advance by mistake and decreased chance of retaining students unnecessarily in the lessons they just finished or the reverse. If a training program must be finished in a hurry, then it is better to set Q so as to minimize the chance of false retainment, P(F-). Thus, Huynh's method gives us more control over the situation, but also brings in more complications of judgement. We don't know how to make the best judgement on the issues, what level a mastery criterion should be set at, and how large the loss ratio Q should be. Neither decisions can be made analytically or in a logical way. Only carefully designed experimental research can answer what are the best decisions.

Let us examine Huynh's method more carefully. Figure 24 shows similar plottings to Figure 23 but the time mastery levels of .70, .75, .85, .90 were also plotted together with .80 on the same screen. The dotted lines were marked by the level of mastery respectively. The horizontal lines correspond to various loss ratios, .50, 1., 1.5,...,20. In Figure 24, the optimal cutoff c_o at the mastery level of .80 is 9 with the loss ratio of 1.00. c_o=9 can be the optimal cutoff at the mastery level of .85 with the loss ratio of Q=2, and also at the mastery level of .90 with Q=2.5. Indeed, the ranges of Q for c_o=9, at 80 percent is from 0 to 1.2, for c_o=9, at 85 percent is from 0.8 to 3, for c_o=9, 90 percent is from 2.25 to 9.25. In the last example, a choice of loss ratio between 2.25 and 9.25 will lead us to select c_o=9 at the mastery level of .90. Figure 24 shows that the range of loss ratio Q for c_o=8 and the mastery level of .90 becomes from 9.25 to over 30. The average loss P(F+) + Q P(F-) associated with Q=9.25 and 30 will be quite different, but P(F-), P(F+) are determined uniquely with c_o=9, and the mastery level θ_o=.9. Test administrators will need more guidance to decide the best loss ratio for their testing.

Table 32

Estimated Number of Misclassified Students

Test	Cutoff[a]		F_+	F_-	Test	Cutoff[a]		F_+	F_-
mvel03	c_0	6	5.3	1.4	mve207	c_0	5	8.5	17.2
	c	7	2.7	5.4		c	4	25.3	4.8
mvel04a	c_0	7	0.2	0.0	mve301	c_0	8	6.9	4.2
	c	10	0.1	0.5		c	8	6.9	4.2
mvel04b	c_0	9	2.9	2.1	mve303	c_0	15	7.6	1.9
	c	9	2.9	2.1		c	16	5.2	4.6
mvel05	c_0	6	2.0	0.8	mve304	c_0	8	3.2	2.0
	c	7	1.0	3.4		c	8	3.2	2.0
mve201a	c_0	7	2.7	0.5	mve305	c_0	5	4.5	0.3
	c	8	1.8	2.0		c	7	1.8	5.7
mve201b	c_0	7	9.3	1.3	mve307	c_0	11	3.3	0.4
	c	8	6.1	4.8		c	12	2.6	1.2
mve202a	c_0	16	11.3	6.1	mve308	c_0	7	6.0	1.2
	c	16	11.3	6.1		c	8	4.1	4.7
mve202b	c_0	5	0.5	0.0	mve401	c_0	7	2.3	0.7
	c	8	0.3	1.0		c	8	1.6	2.8
mve204	c_0	8	8.8	4.4	mve402	c_0	13	14.0	3.7
	c	8	8.8	4.4		c	14	8.6	9.0
mve205a	c_0	8	12.9	12.1	mve403	c_0	8	6.5	2.5
	c	8	12.9	12.1		c	8	6.5	2.5
mve205b	c_0	8	12.4	5.2	mve404	c_0	3	14.1	0.9
	c	8	12.4	5.2		c	4	9.8	5.6
mve206a	c_0	10	3.7	1.4	mve405a	c_0	6	4.0	0.2
	c	11	2.1	4.2		c	8	2.3	3.6
mve206b	c_0	12	4.8	0.9	mve405b	c_0	8	6.9	2.9
	c	14	2.4	7.3		c	8	6.9	2.9
mve206c	c_0	13	0.4	0.0	mve405c	c_0	7	5.5	0.8
	c	16	0.2	0.8		c	8	3.6	3.3

[a] c_0 is the theoretically derived cutoff to minimize $P(F_+) + P(F_-)$.

c is the cutoff actually used in the PLATO Service Program at Chanute.

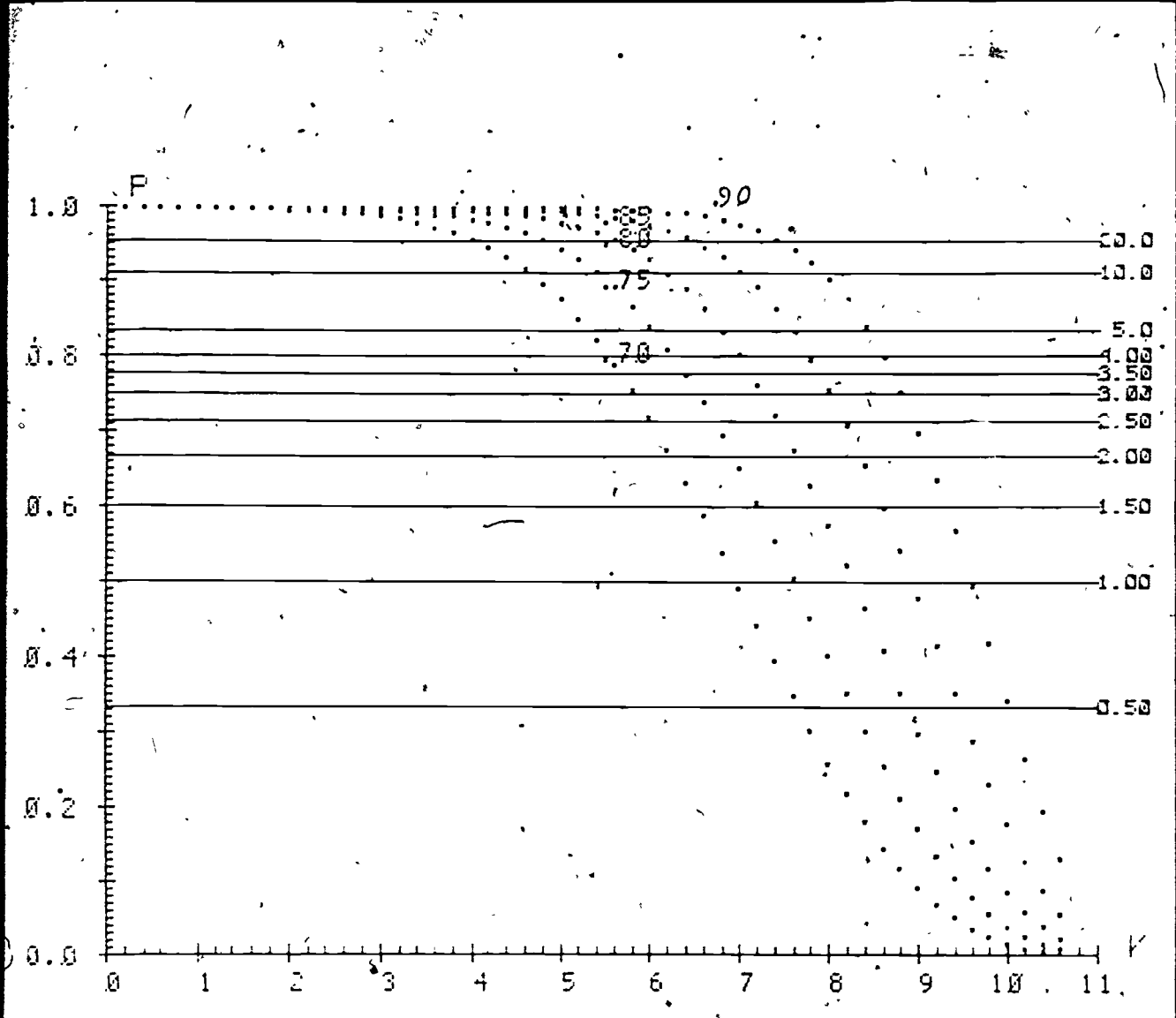


Figure 24 T optimal cutoff for the mastery levels of .70, .75, .80, .85, and .90

lesson = mvel04b subjects = 83 n = 11
mean = 10.1205 SD = 1.7278 α_{21} = 0.8018
a = 2.5012 b = 0.2174

8.3 Other Measures Obtained from the Evaluation Study of the Chanute AFB CBE Project

Correlation Values of Mastery Validation Exams Scores with Block Test Scores and Gain Scores. The evaluation study of the program, supported by the Advanced Research Projects Agency, measured some criterion variables which would be helpful in conducting a validation study of MVEs. The evaluation study revealed that a substantial number of examinees were misclassified (Table 32). Since detailed information on the design used in the evaluation study can be found in Dallmen et al. (1977), just a brief description will be given here.

A 50-item NRT was given at the beginning and the end of the eight-week Chanute Project, which included 35 on-line lessons. The 35 lessons were divided into four subsets called Block1, Block2, Block3, and Block4. After a student studied and mastered all lessons in a block, he took the block test; the block test score was counted in his final grade for the course. He had to take all four block tests, and then a posttest was given in order to measure the effectiveness of the program. Each block test had twenty items which were either multiple-choice or matching. The coefficient alpha reliabilities were not calculated because the tests were written on the PLATO system and the item information was not collected. But α_{21} was available in the following chart. Figure 3 gives a flow chart of the testing program.

In order to validate the effectiveness of lessons, four kinds of correlations were calculated. These correlations are described in the following paragraphs.

Each Block's test scores were matched with the corresponding Master Validation Exam scores and the time needed to master the lesson (mastery time), and their correlations were calculated over the subjects. These two correction values of 27 lessons were denoted by r(B,MVEs) and r(B,time) respectively. Their values are shown in Table 33.

The true gain scores of posttest, x_2, from pretest, x_1, were estimated by multiple regression procedure; the true score difference t_2-t_1 of the observed score difference x_2-x_1 was regressed on the post- and pretest scores. It is known that regression of t_2-t_1 onto the two variables x_1 and x_2 is the same as regressing t_2-t_1 on the scores x_2-x_1 and the residual score, c_2, of x_2 on x_2-x_1 (Tatsuoka, 1975), because the covariance of x_2-x_1 and c_2 equals zero and both x_2-x_1 and c_2 are linear combinations of x_1 and x_2. Therefore, the multiple regression $R(t_2-t_1|x_2-x_1)$ will be given as the sum of the regression of $R(t_2-t_1|x_2-x_1)$ and $R(t_2-t_1|c_2)$:

$$R(t_2-t_1|x_2, x_1) = R(t_2-t_1|x_2-x_1) + R(t_2-t_1|c_2).$$

Table 33

The Correlations of Block tests to MVE Scores and Mastery Time

lesson	r(B, MVEs)	r(B, time)	r(G, MVEs)	r(G, time)
103	.15	-.22	.23	-.38*
104a	.38*	-.33*	.19	-.43*
104b	.36*	.....	.44*	.....
105	.22	-.08	.20	-.34*
201a	.34*	.12	.44*	-.05
201b	.19	-.25	.38*	-.40*
202a	.17	-.04	.07	-.43*
202b	.26	-.03	.28*	-.07
204	.21	-.21	.11	-.13
205a	.28*	-.24	.18	-.32*
205b	.25	-.08	.15	-.26
206a	.40*	-.21	.13	-.22
206b	.12	-.04	-.02	-.18
206c	.00	-.04	.33*	-.08
207	.28*	-.17	.25	-.27
301	.04	-.08	-.11	-.06
303	.34	-.21	.08	-.05
304	.38	-.27	.42*	-.37
305	.07	-.19	.31*	-.26
307	.30*	-.23	.41*	-.30*
308	.01	.04	.00	-.07
401	.50*	-.15	.32*	-.21
402	.25	-.14	.46*	-.34*
403	.40*	-.23	.21	-.02
404	-.02	.00	.02	-.33*
405a	.07	.01	.12	-.11
405b	.25	-.06	.17	-.12
405c	.37*	-.11	.19	-.07

*significant at $p < .05$.

Note that the regression coefficient of the first term is the reliability of gain scores and that of the second term is the increment of multiple R^2. The multiple R is .861, hence the reliability of the multiple regression gain score is R^2=.7405. The squared multiple R of the first term, viz. the reliability of x_2-x_1, is .1047. The squared multiple R of the second term is the increment .6358.

This estimated gain score has a higher reliability than those of pretest and posttest separately. This score was correlated with MVE scores and mastery time. Table 33 shows the result. The numbers of statistically significant correlation values are 12 in Column 2, 1 in Column 3, 10 in Column 4, and 10 in Column 5. The correlation matrix of these four variables over 27 lessons, r(G, MVEs), r(GT, time), r(B, MVEs) and r(B, time) is as follows:

	1	2	3	4
1. r(G, MVEs)	1.000			
2. r(G, time)	-.377	1.000		
3. r(B, MVEs)	.403	-.275	1.000	
4. r(B, time)	-.235	.520	-.468	1.000

Variables 1 and 3 have a moderate correlation value, and Variables 2 and 4 have also a moderate correlation value of .520. The reliability of our gain score has the value of .74 while the four Block tests in Figure 3 have the reliability α_{21} of .56, .33, .47 and .42 respectively which are very low. Therefore, we decided to use only the first two variables, r(G, MVEs) and r(G, time) in subsequent analyses. They were renamed "gain" and "timeg."

The optimal cutoffs c_o that were evaluated in the previous subsection, and designated by c_o in Table 32, were divided by number of items of the corresponding Master Validation Exam.

The distance of c_o from the mean value in each test, $c_o-\bar{x}$, was also divided by the number of items of the corresponding Master Validation Exam in order to make it free from the effect of the test length of MVEs, and then absolute values were taken. This value stands for a sort of the distance of c_o from the mean of each test.

A lesson of Vehicle Training Program at Chanute Air Force Base was said to be validated when 90 percent of the students have achieved the given mastery level of 80 percent of the items answered correctly in the first attempt on each Master Validation Exam. The sample consisted of about 30 students from successive classes. No major modifications of lessons were made until all students in the sample finished the lessons. All lessons were validated according to this criterion between April and September of 1975. These lessons were

used without any major change during the evaluation period and were tested on more students who came in after the validation dates were established. Table 34 includes the information of validation data, the number of examinees who studied the lessons after the lessons were said to be validated (we call this number "nafter" from now on), the percentage of students who achieved the given mastery level at the first try (denoted by % of success), the percentage of students who failed at the first try, the total number of students (which is equal to 30 plus "nafter") and the number of students who passed the end of the lesson test at the first try.

Efficiency Index in the last column in Table 35 (see page 105) is aimed at measuring the quality of Chanute lessons. It is derived from the idea that a good lesson written on the CAI system will allow a student to spend his/her minimum time to master the instructional objective. If a lesson is not good, then a student tends to spend more time than he/she actually needs to master the same instructional goal in a good lesson. The reader might wonder what is the definition of a good lesson. The experienced instructional designer might say that the quality of instruction may be determined by the appropriateness of instructional cues, and the quality and the type of reinforcement given each student, as well as the amount of participation and practice experienced by each student. If the instructional cues are appropriate, clear without ambiguous wording or explanation, then a student must learn the instruction at his/her own learning rate without wasting his/her time.

Carroll (1963), Carroll and Spearitt (1967), and Atkinson (1968), studied the various relationships among the quality of instruction, intelligence and time required for each student to achieve the mastery. Atkinson's findings are especially interesting. They show that students can achieve mastery level of different tasks with different rates and that time variations in learning can be reduced by improving the quality of instruction. Indeed, high quality lessons maximized the individual's learning rate. We all know that a bright student learns very quickly, no matter how poorly a lesson is written. It seems likely that a mediocre student will be the one who suffers the most from ambiguous, unclear instructional cues in a poor quality lesson. If the teaching objective in a lesson does not require previously acquired knowledge or high intelligence, and is fairly easy, then average students should master it as quickly as bright students master it.

How to measure the quality of a lesson became a major concern in the evaluation study of Chanute Air Force Base Computer Based Education Project. They tried to validate a lesson by using success rate (see Tables 31 and 34), but their attempt was not successful. It is natural to consider that the quality of instruction can be recognized at least by two aspects; one is higher success rate, the other is faster learning rate.

Table 34

Summary of Master Validation Exams in the Chanute PLATO IV Project

Lessons	M[a]	Validation Date	Size of tested out sample	% of Success	% of Failure	Total N	# of Success
103	30	10 June	63	89%	11%	93	83
104a	30	14 April	114	94%	6%	144	134
104b	30	14 April	113	86%	14%	143	124
105	30	14 April	102	88%	12%	132	117
106	30	19 June	33	82%	18%	63	54
201a	30	28 May	99	90%	10%	129	116
201b	30	23 May	109	72%	28%	139	105
202a	30	18 Aug	33	82%	18%	63	54
202b	30	28 May	90	98%	2%	120	115
203a	30	28 May	33	97%	3%	63	59
203b	30	13 June	33	94%	6%	63	58
203c	30	18 Aug	33	91%	9%	63	57
204	30	18 Aug	33	94%	6%	63	58
205a	30	15 Jan	33	79%	21%	63	53
205b	30	15 Jan	33	82%	18%	63	54
206a	30	13 June	90	82%	18%	120	101
206b	30	25 June	65	82%	18%	95	80
206c	30	11 April	118	95%	5%	148	139
207	30	15 Aug	33	91%	9%	63	57
301	30	25 June	109	79%	21%	139	113
304	30	25 June	65	82%	18%	95	80
305	30	18 May	109	96%	4%	139	132
307	30	14 April	130	81%	19%	160	132
308	30	18 May	109	63%	37%	139	96
401	30	17 April	142	83%	17%	172	146
402	30	8 July	65	79%	21%	95	78
403	30	30 June	65	79%	21%	95	78
404	30	2 Sept	33	100%	0%	63	60

[a]M is the sample size used for establishing validation dates.

(Table 34 cont.)

Lessons	M[a]	Validation Date	Size of tested out sample	%% of Success	% of Failure	Total N	# of Success
405a	30	26 Aug	33	100%	0%	63	60
405b	30	26 Aug	33	91%	9%	63	57
405c	30	26 Aug	33	94%	6%	63	58
405d	30	2 Sept	33	73%	27%	63	51
406	30	30 June	65	95%	5%	95	89
407	30	22 Sept	33	88%	12%	63	56

Tatsuoka (1978) discussed the possibility of using the success rate as a measure of instructional quality in her paper, and the result was not favorable. Success rate measure depends on the scores on the end-of-lesson test, a criterion-referenced test which has been a problem in educational measurement. It is dangerous to use a criterion-referenced test alone as a measure of the instructional quality, and the success rate is contaminated by the problems of misclassifications, false positive, and false negative. It is urgent to establish a method that can measure the quality of instruction directly without using criterion-referenced testing as an auxiliary means. We believe our efficiency index provides one such wanted measure. The procedure for deriving the efficiency index is as follows.

1. The total sample of about 80 subjects was divided into three groups according to their scores on the aptitude test, the Armed Services Vocational Aptitude Battery (ASVAB). The test is aimed at measuring general-technical, mehcanical, motor mechanical and electronics aptitudes for high school seniors, as part of the recruiting programs of the Army, Navy, and Air Force. The first group consists of the top 25 percent of the students, the second is the middle 50 percent of students and the third is the bottom 25 percent of the students who took the ASVAB. The average mastery times of the three groups are calculated and summarized in Table 35. The t-test of mean mastery times for the two groups, Group 1 and Group 2, revealed that 9 out of 27 lessons were statistically significant at $p < .05$.

2. Lesson MVE201a was arbitrarily picked as the base, and its mean mastery times in Groups 1 and 2 were divided by the respective mean values of mastery time of every other lesson. We calculated such ratios of the mastery time of 27 Chanute lessons in Groups 1 and 2, taking the mean mastery of lesson MVE201a as the base.

3. According to the assumption that a good lesson will not make the average students slow down to master it in comparison with time taken to master the lesson for the brighter students, we divided the newly calculated 27 ratios, [mean mastery time of MVE201a]/[mean mastery time of lesson X], in Group 1 by the corresponding ratios in Group 2, and obtained 27 efficiency indices which appear in the last column of Table 35.

If the value of efficiency index of lesson A is larger than that of lesson B, then we might be able to say that lesson A is more efficient than lesson B.

8.4 The Results of Statistical Analyses Over 27 Chanute Lessons

Nineteen measures were selected and their correlation matrix was calculated. Table 36 gives a brief description of 19 variables and Table 37 is the correlation matrix of these variables.

Table 35

Average Mastery Time and Efficiency Index

Lesson	Mean and Standard Deviation (Minutes) 1*	2**	3***	Efficiency Index
MVE103	21.25 , 5.26	32.95 , 15.42	43.83 , 27.11	0.746
MVE104a	23.42 , 5.09	36.82 , 13.61	36.25 , 13.07	0.736
MVE105	31.73 , 8.19	41.63 , 12.32	54.42 , 23.00	0.882
MVE201a	11.20 , 5.35	12.96 , 6.31	13.75 , 7.19	1.000
MVE201b	27.08 , 16.02	42.46 , 23.05	52.42 , 29.05	0.738
MVE202a	142.23 , 56.22	183.44 , 73.65	218.14 ,114.81	0.897
MVE202b	12.46 , 3.89	14.58 , 4.46	14.25 , 3.41	0.989
MVE204	71.64 , 31.36	100.76 , 59.91	102.50 , 60.03	0.823
MVE205a	86.75 , 25.32	111.60 , 47.97	149.17 , 94.90	0.899
MVE205b	27.90 , 11.80	44.46 , 35.94	50.18 , 29.35	0.726
MVE206a	37.89 , 12.75	53.10 , 21.77	55.00 , 26.03	0.825
MVE206b	11.00 , 3.20	20.33 , 13.50	22.50 , 7.15	0.626
MVE206c	33.13 , 15.50	50.95 , 33.75	41.78 , 10.69	0.752
MVE207	22.45 , 5.05	34.50 , 16.18	43.15 , 26.86	0.753
MVE301	26.67 , 9.10	29.81 , 18.81	29.92 , 16.98	1.035
MVE303	11.57 , 3.99	13.50 , 7.06	15.36 , 6.99	0.992
MVE304	12.83 , 7.25	10.06 , 4.99	15.80 , 7.71	1.476
MVE305	14.75 , 3.41	19.90 , 9.01	21.80 , 7.15	0.858
MVE307	44.00 , 12.54	58.22 , 27.57	86.83 , 22.74	0.874
MVE308	38.00 , 5.95	44.71 , 18.66	42.10 , 15.77	0.983
MVE401	17.00 , 3.27	21.06 , 5.50	26.17 , 5.42	0.934
MVE402	53.55 , 18.91	81.69 , 67.17	114.08 , 49.75	0.758
MVE403	7.13 , 1.13	12.88 , 16.34	15.86 , 8.73	0.666
MVE404	10.20 , 6.14	10.00 , 5.13	13.44 , 7.50	1.180
MVE405a	23.00 , 14.10	25.37 , 7.21	32.60 , 13.82	1.049
MVE405b	33.25 , 9.16	42.47 , 18.37	39.11 , 19.33	0.906
MVE405c	9.00 , 2.38	11.10 , 8.51	13.00 , 5.29	0.938

*The top 25 percent of examinees according to ASVAB scores.

**The middle 50 percent of examinees according to ASVAB scores.

**The bottom 25 percent of examinees according to ASVAB scores.

Table 36

A Brief Description of 19 Variables

Variable Number	Notation	Description
1	P(F+)	The probability of false positive
2	c_o/n	The optimum cutoff ratio so as to minimize misclassifications
3	α_{21}	The ratio of true variance to observed variance
4	P(F+) + P(F-)	The probability of misclassification
5	nafter	The number of subjects tested after a lesson was declared to be validated
6	% fail	Observed percentage of failure in MVE
7	t_o	The minimum time parameter from Weibull distribution
8	mc	Maximum correlation from estimation procedure of Weibull parameters
9	c	Shape parameter of Weibull distribution
10	μ_o	Scale parameter of Weibull distribution
11	p	Probability value from Kolmogorov-Smirnov test
12	range	Maximum mastery time minus minimum mastery time
13	efficiency index	Relative ratio of mean mastery time of higher aptitude group to mediocre aptitude group
14	gain	Correlation of gain scores with MVE scores
15	time g	Correlation of gain scores with mastery time
16	items	Number of items in a test
17	$\lvert c_o-\text{mean}\rvert/n$	Relative distance of c_o from the mean
18	P(F-)	Probability of false negative
19	P(A or F+)	Probability of pass based on the observed cutoff c

The probability of false positive (or advancement), P(F+) has correlation values of .931, -.562, -.678, .638 and -.637 with P(F+) + P(F-), nafter, $\frac{|c_o-\text{mean}|}{n}$, P(F-) and P(A or F+) respectively. According to these correlations, when false positive occurs, then false negative more likely occurs but the observed passing rate, P(A or F+) more likely declines. That means that the lessons whose

Table 37

Correlation Matrix of 19 Variables
(x 1000)

	1	2	3	4	5	6	7	8	9	10
1	1000	250	-6	931	-562	111	147	-345	-276	342
2	250	1000	358	393	-373	167	754	259	-10	614
3	-6	358	1000	-20	-37	384	158	44	-236	196
4	931	393	-20	1000	-617	165	257	-223	-294	396
5	-562	-373	-37	-617	1000	335	-261	285	337	-329
6	111	167	384	165	335	1000	225	1	-20	254
7	147	754	158	257	-261	225	1000	203	20	762
8	-345	259	44	-223	285	1	203	1000	389	32
9	-276	-10	-236	-294	337	-20	20	389	1000	60
10	342	614	196	396	-329	254	762	32	60	1000
11	182	270	89	279	-64	76	318	295	-31	383
12	265	621	213	345	-304	206	755	67	-15	967
13	48	19	-145	-1	-30	-238	-167	-101	401	-227
14	-283	-244	90	-264	271	32	-90	-205	64	-94
15	183	-233	-259	54	-99	-460	-508	-26	-71	-449
16	-108	-271	172	-211	426	385	-106	98	43	43
17	-678	-441	-626	-662	353	-522	-219	239	228	-440
18	638	542	79	869	-544	293	376	-17	-234	428
19	-637	-558	-189	-853	587	-289	-426	12	298	-473

	11	12	13	14	15	16	17	18	19
11	1000	430	-193	-126	-175	157	-259	347	-326
12	430	1000	-289	-74	-414	70	-402	417	-467
13	-193	-289	1000	-1	261	-319	105	-75	148
14	-126	-74	-1	1000	-377	231	181	-196	155
15	-175	-414	261	-377	1000	-190	119	-171	174
16	157	70	-319	231	-190	1000	-123	-264	238
17	-259	-402	105	181	119	-123	1000	-595	640
18	347	417	-75	-196	-171	-264	-595	1000	-974
19	-326	-467	148	155	174	238	640	-974	1000

observed passing rate, P(A or F+) is higher tend to have less chance of false positive (advancement) cases. The test which advances the students to the next lesson more frequently by mistake tends also to retain the student whose true scores are really above the mastery level. The high correlation of P(F+) and $\frac{|c_o-\text{mean}|}{n}$ shows when the observed cutoff c_o is closer to mean, then the misclassification of false advancement tends to occur more often. The correlation value of -.562 with the variable, nafter, the number of students who studied a lesson after the validation date was set (if over 90 percent of the students pass the mastery level of a MVE, then the lesson was said to be validated) indicates that the probability P(F+) will be small if the lessons whose validation date were established at an earlier date during the period of evaluation study at PLATO program.

This relation is true for the variables P(F+ or F-) and P(F-) because the correlations of variable "nafter" with them are -.617 and -.544 respectively. Moreover, P(F+), P(F-) and P(F+ or F-) correlate highly with variable $\frac{|c_o-\text{mean}|}{n}$ with the values of -.678, -.595, and -.662 respectively. But the correlations between "nafter" and $\frac{|c_o-\text{mean}|}{n}$ is not so low, at .353. Further discussion of the appropriateness of the procedure that a lesson can be said validated will be found in Tatsuoka (1978).

Variable 17 ($\frac{|c_o-\text{mean}|}{n}$) correlates significantly with nine variables, and so does Variable 19 (P(A or F+)). Variable 12, (range) and Variable 18 (P(F-)) each correlate significantly with eight variables. Variable 2 (c_o) has seven variables, Variables 4 (P(F+ or F-) and 10 (μ_o) have six variables whose correlation values are significant. In order to clarify the characteristics of the 19 variables, principal component analysis was first performed. The first five eigenvalues were 6.32, 3.08, 2.12, 1.40, 1.30 respectively and their cumulative percentage was 75 percent of the total variance. The factor matrix was orthogonally rotated by Varimax analysis and five factors were selected. Table 38 summarizes variables in each factor with their factor loadings.

Table 38

The Results of Factor Analysis

Factor 1			Factor 2		
Variable		Loading	Variable		Loading
1	P(F+)	.89	2	c_o/n	.70
4	P(F+ or F-)	.94	7	t_o	.90
5	nafter	-.69	10	μ_o	.85
17	$\lvert c_o$-mean$\rvert/n$	-.73	12	range	.85
18	P(F-)	.81	15	timeg	-.66
19	P(A or F+)	-.81			

Factor 3			Factor 4		
Variable		Loading	Variable		Loading
3	α_{21}	.62	8	mc	-.78
6	% fail	.85	11	p	-.54
16	items	.62	14	gain	.61
17	$\lvert c_o$-mean$\rvert/n$	-.58			

Factor 5		
Variable		Loading
9	c	.69
13	efficiency index	.84

Probability Variables 1, 4, 18 and 19 clustered together with Variables 5 and 17 as Factor 1. Time variables 7, 10, 12 and 15 clustered together with the optimal cutoff c_o. The result most interesting to the authors was Factor 5, the shape parameter of Weibull distribution clustering together with the efficiency index of lessons. The correlation of c and efficiency index is .401 which means that if c is larger, then the lessons tend to have larger efficiency index, and hence the difference between the average mean time of Group 1

and Group 2 becomes smaller, with respect to the difference of those in Group 1 and 2 of lesson MVE201a. That means by our assumption that if c is larger, then the corresponding lesson is more efficiently teaching students. Recall the previously developed argument that CRR of larger than 1 was interpreted as meaning that students engaged themselves with the task of solving a problem, and CRR of smaller than 1 indicated they gave up a given item because it was too difficult to try for them. These two results from the analysis of lessons and test items were independently derived in different contexts, and yet both make sense and sound reasonable. Since aptitude scores are infrequently available in common practice, it is usually difficult to obtain the efficiency index we introduced in this report. But mastery time can be obtained fairly easily from lessons written on a CAI system, so our research result will be used to measure some aspect of quality in lessons, we hope.

Multiple Regression Analyses were performed in the several sets of variables. The purpose of the analysis was to see which variables predict large misclassifications. Variables 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16 and 17 were taken as a set of predictors, and Variable 1 (P(F+)) was taken as the criterion. Stepwise multiple regression where F values of entry and removal of predictors were set at 2.0 was performed on these variables, and then $|c_o\text{-mean}|/n$ with t-value of -10.4, α_{21} with t = -6.6, % fail with t = 2.6 and c with t = -2.0 were selected to predict P(F+), probability of false positive. Multiple R was .921; R corrected for shrinkage was .828.

A second analysis was performed on the same set of predictors and the criterion variable of 18, P(F-). Multiple R of .912, R corrected for shrinkage of .782 were obtained with the predictors $|c_o\text{-mean}|/n$, α_{21}, maximum correlation, number of items, c, and timeg (correlation of gain and MVE scores).

A third analysis was done on the same set of predictors and the criterion variable 4, P(F+) + P(F-). The result was pretty much similar to the first and second analysis results. The predictors are $|c_o\text{-mean}|/n$, α_{21}, maximum correlation, c, the correlation of gain and MVE scores, nafter, and number of items in a test. Multiple R is .970 and R corrected for shrinkage is .918. These results are summarized in Tables 39, 40 and 41.

Table 39

Relationship between P(F+) and Other Variables

Variable		Beta-Coefficient	SDT Error	t
3	α_{21}	-.709	.108	-6.6
6	% fail	-.256	.099	-2.6
9	c	-.172	.087	-2.0
17	$\lvert c_o\text{-mean}\rvert/n$	-1.216	.117	-10.4

Multiple R = .921, Corrected R for shrinkage = .828, $F_{4,22} = 30.543$

Table 40

Relationship between P(F-) and Other Variables

Variable		Beta-Coefficient	SDT Error	t
3	α_{21}	-.793	.141	-5.6
8	mc	.529	.123	4.3
9	c	-.368	.110	-3.4
15	timeg	-.238	.106	-2.2
16	items	-.409	.101	-4.1
17	$\lvert c_o\text{-mean}\rvert/n$	-1.198	.145	-8.3

Multiple R = .912, Corrected R for shrinkage = .782, $F_{7,19} = 13.446$

Table 41

Relationship between P(F_+ or F_-) and Other Variables

Variable		Beta-Coefficient	SDT Error	t
3	α_{21}	-.864	.088	-9.8
5	nafter	-.195	.077	-2.5
8	mc	.377	.079	4.8
9	c	-.362	.080	-4.5
14	gain	.224	.070	3.2
16	items	-.161	.073	-2.2
17	$\lvert c_o\text{-mean}\rvert/n$	-1.216	.096	-12.6

Multiple R = .970, Corrected R for shrinkage = .918, $F_{8,18} = 35.382$

Variable $|c_o\text{-mean}|/n$ is a common predictor of three criteria variables and t-values are -10.4, -12.6 and -8.3 which are the largest among other predictors. This result is expected due to the nature of beta-binomial model, but α_{21} as the second strongest predictor in the three analyses is surprising. If α_{21} is high enough, then the probability of the three errors, false positive, false negative and either misclassification, will be minimized. Most Master Validation Exams have reliabilities of around .4 to .5 which is quite low, so it is natural to expect that misclassifications will have occurred quite frequently in the program.

The variable α_{21} does not correlate significantly with Variable 16, number of items in the tests; it correlates with Variable 6, percentage of failure at the 5 percent significance level. This relationship may be interesting to investigate further, especially when the test lengths are short and about the same, 10 to 15 items as is typical for criterion referenced tests. It is apparent that α_{21} is a strong predictor of the three criteria with beta values of -.709, -.865, and -.793 respectively, and therefore internal consistency is an important factor for controlling the occurrences of misclassifications in a criterion-referenced test. Figure 25 is a copy of the PLATO screen where the graphic relationship between P(F+) + P(F-) and α_{21} was plotted. The curves in Figure 25 are of P(F+) + P(F-) as y-axis, α_{21} as x-axis for the test whose mean value is 8.907 and the test length is 10. When cutoffs are 7, 8, and 10, the corresponding curves go down as α_{21} goes larger. The curve for cutoff 6 has the optimum value at around α_{21} = .6, but it goes down as α_{21} increases. If internal consistency α_{21} of the test is between .53 and 1, then cutoff 7 minimizes the probability of F+ or F-. If α_{21} of the test is less than .53, then the optimum cutoff will be 6. Thus, the optimum cutoff scores so as to minimize the misclassification mistakes depends on α_{21}. This fact will be one useful guide to construct a criterion-referenced test so that misclassifications, false positive, false negative can be minimized. The most interesting result is that the shape parameter c appears in three cases as a predictor with beta-values of -.172, -.362 and -.368 respectively. If the lesson has larger c value, then the probabilities of misclassification, false positive, and false negative become smaller. Even though P(F+), P(F-) and P(F+) + P(F-) are determined by such variables as number of items, means of CR test scores, α_{21} that are purely obtained only from a test, the value of the shape parameter c of the Weibull distribution entered as a common predictor of the three misclassification cases. It implies that some factor of a lesson related to the quality of lessons, or conditional mastery rate of the lesson (conditional probability of a student who has not mastered the lesson at time t will master it at the next moment, $t + \Delta t$) affects the possibility of having misclassifications upon judging based on the scores on the end of lesson test.

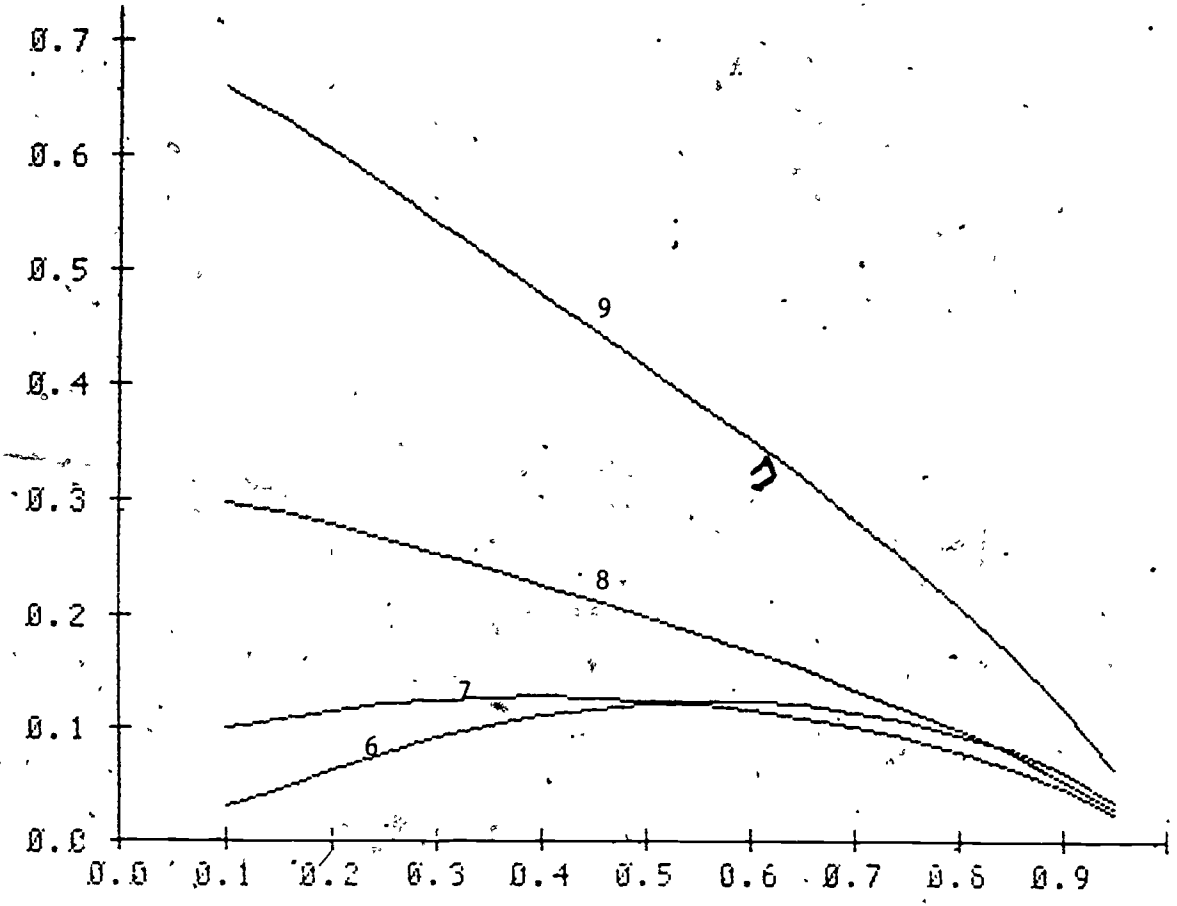


Figure 25 The relation between $P(F_+)+P(F_-)$ and α_{21}
mean = 8.907, no. of items = 10, cutoffs 6,9

In the last analysis, Variable 2, c_o/n, was taken as the criterion and all other variables as predictors. Stepwise regression analysis selected predictors, t_o, $|c_o\text{-mean}|/n$, items and mc with multiple R of .875 and R corrected for shrinkage of .735; F value for this regression was 17.98. Table 42 shows the result of analysis.

Table 42

Relationship between c_o and Other Variables

Variable		Beta-Coefficient	Standard Error	t
7	t_o	.578	.112	5.2
8	mc	.269	.112	2.4
16	items	-.287	.107	-2.7
17	$\|c_o\text{-mean}\|/n$	-.414	.113	-3.7

Multiple R = .875, Corrected R for shrinkage = .735, $F_{4,22} = 17.984$

It is surprising to see that t_o, location parameter of the Weibull distribution, enters as the strongest predictor of the optimum cutoff scores c_o with t-value of 5.2. Beta-coefficient of .578 indicates the lessons that have larger t_o values tend to have larger c_o/n. Note that c_o came in together with t_o, μ_o, range and timeg. The percentage score of the optimum cutoff, $|c_o|/n$ showed a strong relationship with time variables in the corresponding lesson. We don't know how to interpret this result.

The major conclusion of this section is that misclassification, false positive and false negative are mainly affected by three factors: how closely to the mean of a test the cutoff was selected, internal consistency of a test, and conditional mastery rate of a lesson.

9. SUMMARY AND CONCLUSIONS

This study investigated the feasibility of using the family of Weibull distributions--a family which is widely used in system-reliability analysis--as a model for the distributions of time scores (response times) of items in criterion-referenced tests, lesson segments and entire lessons that were implemented on the PLATO system. The items were those of a series of matrix algebra tests developed for the dual purpose of using in this study and for testing students in three statistics courses at UIUC both before and after they studied our matrix algebra course. The latter provided the lesson segments (including exercises), while the entire lessons came from the Chanute AFB CBE Project and deals with special and general vehicle maintenance training.

The fits of the Weibull distributions to these various observed distributions were, on the whole, very good to excellent as gauged by the Kolmogorov-Smirnov goodness-of-fit test. However, for some items (most of which possessed certain exceptional properties in common) the two-parameter gamma distribution offered better fits. The same held true with even greater force for the exercises occurring in the matrix algebra lessons. Tentative explanations of when and why the gamma was better than the Weibull were advanced, but discovery of definitive reasons must await future research.

Interpretations of the three Weibull parameters--the theoretical minimum time or location parameter t_0, the scale parameter μ_0 which is closely related to the mean, and the most interesting although sometimes "recalcitrant" shape parameter c--were given in terms of psychometric properties of the achievement test items. The last mentioned parameter was found by correlational analysis to be moderately related to two kinds of item difficulty index--the traditional proportion passing and a more subtle one developed very recently by Irmgard Loeschner (personal communication). It was also believed to be related to what might be called "degree of engagement or involvement" of the student with the task, and further to be associated with degree of familiarity with it. Both these are akin to, but conceptually different from, difficulty.

A function related to, and partially determined by, the shape parameter c is what we dubbed the conditional response rate (CRR) and which is called the hazard rate in the system-reliability literature. This is the conditional probability that an examinee who has not responded to an item (or lesson segment, etc.) up to time t will respond to it within the following infiniteimal interval $[t, t + \Delta t]$. When $c < 1$ CRR is a monotonically decreasing function of t, and the implication is that students give up early trying to solve such an item. This typically occurs in pretest items, while the same items given after the instruction usually has $c > 1$ and the CRR is a monotonically increasing curve. However, anomalous items (of which there

were three in one of our subtests) that involve material not covered in the lesson will behave like pretest items even when given in the posttest. In a way, therefore, one might say that the Weibull shape parameter c relates also to how well an item "matches" the instructional content. If the match is poor (as it was in the three anomalous items), then the students will get frustrated and angry (as they did) and will quit trying early, which will be reflected in c becoming less than 1. If the match is good, on the other hand, the students will by and large become ego involved and will engage themselves deeply with the items, thus resulting in c > 1 which leads to an increasing CRR, implying that the longer a student perseveres in the item the greater the chances that he/she will answer it.

A relatively trivial point, but nonetheless one which bears passing mention, is the fact that the location parameter t_0 estimated for the group which got that item right (i.e. the "OK subgroup" as we have been calling it) gives a good idea of the minimum time that should be allotted for answering that item.

Another finding is that the time-score distribution of an item which requires only simple, mechanical subtasks for its execution is generally fitted better by a two-parameter gamma than by a Weibull distribution. As mentioned in Section 2.2, a two-parameter gamma distribution [see equation (2.9), p. 10] with integer-valued c(>1) is a c-fold convolution of one-parameter negative exponential distributions. Such distributions fit well the time distribution of a simple task with but one stage; hence their c-fold convolution should fit a problem consisting of c independent stages each of which is simple and mechanical. Thus the finding just cited makes good, intuitive sense.

We also found some evidence to support the thesis that the shape parameter c is a more sensitive measure of the "conceptual difficulty" of an item than is the traditional difficulty index. This was done by identifying five sets of items that respectively had the same difficulty in the traditional sense but differed considerably in their c values. For example, both the following items were correctly answered by 29 percent of our sample: (1) If AB = AC, then is B = C? (2) An item calling for the inverse of a 2x2 matrix. Yet c = 1.01 for the first and c = 1.24 for the second, and certainly it can be argued that the latter is conceptually more difficult than the former.

A new measure which we named the "efficiency index" of a lesson was defined as follows. The total sample of students is divided into three groups on the basis of scores on an aptitude test relevant to the subject matter of the lessons (say A and B) whose relative efficiencies or qualities are to be compared. The groups are, for instance, the top 25 percent (group 1), the middle 50 percent (group 2) and the bottom 25 percent (which is discarded from further consideration). We assume that there are other lessons in the same or similar subject matter that have also been studied by our sample

of students, and one of them is arbitrarily chosen as a "reference lesson" (R). The average times taken by group 1 and group 2 to master the reference lesson are divided respectively by the mean mastery times of Lesson A and Lesson B in two groups. We now have four ratios,

$$\bar{X}_{R1}/\bar{X}_{A1},\ \bar{X}_{R2}/\bar{X}_{A2},\ \bar{X}_{R1}/\bar{X}_{B1} \text{ and } \bar{X}_{R2}/\bar{X}_{B2}, \text{ say.}$$

Finally, we take the pairwise ratios of these ratios, thus:

$$E_{A(R)} = \frac{\bar{X}_{R1}/\bar{X}_{A1}}{\bar{X}_{R2}/\bar{X}_{A2}} \text{ and } E_{B(R)} = \frac{\bar{X}_{R1}/\bar{X}_{B1}}{\bar{X}_{R2}/\bar{X}_{B2}}.$$

On the reasonable assumption that a "good" lesson will not require group 2 (average aptitude) students much more time than group 1 (high aptitude) students to master it, while a "poor" lesson will show a larger discrepancy in mastery times, the ratios $E_{A(R)}$ and $E_{B(R)}$ defined above will represent the relative efficiencies of lessons A and B: the one with the larger the ratio is the more efficient lesson. If there are more lessons to be compared, there will be more such efficiency indices, and the lessons will be rank ordered by them. (The rank ordering will be invariant of what lesson is chosen as the reference lesson.)

When a factor analysis followed by varimax rotation was carried out on 19 variables including our efficiency index and the Weibull shape parameter c, a distinct factor was found that loaded only these two variables. We thus find yet another evidence of the meaningfulness of parameter c.

The relationship between the probability P(F+) of a false positive (calling a non-master a master on the basis of a criterion-referenced test), the probability P(F-) of a false negative (calling a master a non-master) and the probability P(F+ or F-) of either misclassification on the one hand, and the three Weibull parameters, other psychometric properties of tests such as α_{21} and $|c_o - \text{mean}|/n$ (the distance between the mean and the theoretical cutoff point for declaring "masterhood," adjusted for test length) was examined by stepwise multiple regression analysis. It turned out that the shape parameter c was one of the strongest predictors of P(F+) and of P(F-), along with α_{21} and $|c_o - \text{mean}|/n$. The direction of the relationship so far as c is concerned was that, the larger the c the smaller the P(F+) and P(F-). (Actually, the same directionality of relationship held for α_{21} and $|c_o - \text{mean}|/n$ as well.) Hence we may

conclude that, although one way to minimize misclassifications is naturally to use the optimal cutoff point, that alone is insufficient. We may still have quite large P(F+), and P(F-) and P(F+ or F-) values for some tests unless internal consistency (α_{21}) and c (a surrogate measure of efficiency of instruction) are also high.

One incidental but in our mind important and interesting finding was that item discrimination power appears to be an "inverted-U" type function of time allowed for completing that item. This is how we arrived at this conclusion.

Carroll (1963) emphasized in his "model of school learning" the importance of differences in the time required to learn and asserted that learning rate was an important source of individual differences in educability. A study conducted by one of the present authors during the past year showed that the time needed to complete certain tests correlates with aptitude scores more significantly than do the scores on the tests. Sato and his coworkers (1973, 1975), and Tatsuoka and Tatsuoka (1978) have studied the statistical aspects of time-score distributions and their characteristics. When a test item is easy, there is an optimal time point within a relatively short time interval such that the discriminating power of the item becomes the largest. On the other hand, for difficult items, the longer the time allowed the better the discriminating power. Figure 10 (p.41) is a copy of the PLATO screen display of plots of the discriminating powers of an item in our matrix algebra test, against 10 time points obtained in the following manner. The subjects were first arranged in ascending order of the time they took to respond to a given item. The first (leftmost) point in the figure was obtained as follows. Only those who got the item right and were in the fastest 10 percent of the group were given a score of 1. Everyone in the remaining 90 percent of the group got a score of 0 even if they got the item right. The point-biserial correlation coefficient calculated between the item score thus defined and the modified total score is the ordinate of the first point (10, .02) in Figure 10. Next only those who got the item right and were in the fastest 20 percent were scored 1, and the others were scored 0 on the item, and the total score was accordingly modified. The point-biserial correlation thus calculated is the ordinate of the second plotted point (20, .14). The same process was repeated for the remaining cutoff percentages, 30 percent, 40 percent, ..., 90 percent, yielding adjusted discriminating powers, .27, .46, ..., .15 respectively.

The limitations of this study are many in number, perhaps the chief of which is the fact that it is not experimental in the sense of having a neat design and experimenter-manipulated independent variables. It is, rather, a status study from which, of course, causal relations cannot be definitively concluded but only inferred and hinted at. On the other hand it has the strength of having been conducted

in a real CAI classroom situation yielding "dirty" data instead of "antiseptic" data that often accrue from tightly controlled laboratory experiments which are hence frequently criticized as bearing little relationship with real life. (To be sure, some of the dirty data were "laundered" to the extent that they meet the minimal demands of analyzability--not to fit our preconceived theory of course--but dirty and "real-lifeish" they nevertheless remained.)

Other weaknesses, as mentioned in the main text, were (1) that the parameter-estimation procedures were not the best conceivable or even available--we learned too late of the best existing method, via. an iterative maximum-likelihood approach; and (2) that we did not consider two-component composite Weibull distributions which probably would have fit the total sample without our having to partition it into the "OK" and "NO" subgroups--those who answered an item (or exercise, etc.) right or wrong, respectively.

As of this writing we have in fact implemented on the PLATO system a program for the iterative maximum-likelihood method (adapted from the FORTRAN printout kindly supplied to us gratis by Dr. H. Leon Harter of the Wright-Patterson AFB, Ohio) which, mutatis mutandis, is usable for estimating the parameters of both the three-parameter gamma and the Weibull distributions in the best possible way given the state of the art. We intend to do this as well as experiment with composite Weibull distributions in the near future.

Thus, we would be the first to concede that we have barely scraped the surface in studying the utility of response time (time scores) along with performance scores for analyzing and evaluating data from criterion-referenced tests, both for the purpose of assessing the quality of the tests themselves and for improved testing of the examinees' abilities.

Nevertheless, we believe that we have at least demonstrated the feasibility of this approach and hope to have shown that further research along these lines is warranted. In particular, the Weibull distribution in its two-parameter form (which we used in this study), three-parameter form, or two-component composite form--long used by system-reliability analysts but apparently not widely known among educational and psychological researchers--seems to bear further investigation for this purpose.

REFERENCES

Atkinson, R.C., Computer-based instruction in initial reading. In proceedings of the 1967 invitational conference on testing problems. Princeton, Educational Testing Service, 1968, 58-67.

Bargman, R.E., Association in a class of growth functions. Urbana, University of Illinois, 1966.

Block, J.H., (Ed.) Mastery learning: theory and practice. New York: Holt, Reinhart & Winston, 1971.

Brée, D. S. The distribution of problem-solving times: an examination of the stage model. British Journal of Mathematical and Statistical Psychology, 1975, 28, 177-200.

Carroll, J.B., A model of school learning. Teachers College Records, 1963, 64, 723-733.

Carroll, J. B., & Spearitt, D., A study of a model of school learning. Monograph No.4 Cambridge, Massachusetts: Harvard University, Center for Research and Development of Educational Differences, 1967.

Emrick, J. A. An Evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Guttman, L., The quantification of a class of attributes: a theory and method of scale construction. In P. Horst(Ed.), The prediction of personnal adjustment. Social Science Research Council, Bulletin 48, 1941, 321-345.

Harter, H.L. & Moore, A.H., Maximum likelihood estimation of the parameters of the Gamma and Weibull populations from complete and from censored samples, Technometrics, 1965, 7, 639-643.

Harris, C.W., An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of educational measurement. 1972, 9, 27-29.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 43-64.

John, M.V., Jr. & Lieberman, G.J., An exact asymptotically Efficient confidence bound for reliability in the case of the Weibull distribution, technometrics, 1966, 8, 135-175.

Keats, J.A. & Lord, F.M., A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72

Lennon, G.H., Maximum-likelyhood estimation for the three parameter Weibull distribution based on censored-sample, Technometrics, (to appear)

Linn, R.L. Personal communication, October 1, 1978.

Livingston, S.A., Criterion-referenced applications of classical test theory. Journal of educational measurement, 1972, 9, 13-25.

Loeschner, I., Personal communication, February 11, 1978.

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

Mann, N.R., Tables for obtaining the best linear invariant estimates of parameters of Weibull distribution, Technometrics, 1967, 9, 629-645.

Mann, N.R., Optimum estimators for linear function of location and scale parameters, Annals of mathematical statistics, 1969, 40, 2149-2155.

Mann, N. R., Schafer, R. E. & Singpurwalla, N. D., Methods for Statistical analysis of reliability and life data. John Wiley & Sons, New York, 1974.

Millman, J. Tables for determining number of items needed on domain-referenced tests and number of students to be tested. Los Angeles: Instructional Objectives Exchange, Technical Paper No.5, April 1972.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-215.

Novick, M.R., The axioms and principal results of classical test theory. Journal of mathematical psychology. 1966, 3, 1-18. 1961, 4, 321-324.

Novick, M. R. & Lewis, C. Prescribing test length for criterion-referenced measurement. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angels: UCLA Graduate school of Education, Center for the Study of Evaluation, 1974.

Appendix A

Sample Pages of Matrix

Algebra Lessons

The numerical entries are called the elements of the matrix.
A particular element is specified by the number of the ROW and the number of the COLUMN in which it occurs.

Here is a 4×3 matrix
3 columns

4 rows

$$\begin{bmatrix} 2 & 4 & 7 \\ 3 & 8 & 5 \\ 1 & 6 & 9 \\ 8 & 6 & 2 \end{bmatrix}$$

What is 1st row vector?
↓1st ↓2nd ↓3rd-elements
[, ,].

Press NEXT to continue, SHIFT-HELP for index
BACK to see previous page

Appendix A

Sample Pages of Matrix

Algebra Lessons

The numerical entries are called the elements of the matrix.
A particular element is specified by the number of the ROW and the number of the COLUMN in which it occurs.

Here is a 4×3 matrix
3 columns

4 rows

$$\begin{bmatrix} 2 & 4 & 7 \\ 3 & 8 & 5 \\ 1 & 6 & 9 \\ 8 & 6 & 2 \end{bmatrix}$$

What is 1st row vector?
↓1st ↓2nd ↓3rd-elements
[, ,].

Press NEXT to continue, SHIFT-HELP for index
BACK to see previous page

3.3 Evaluation of the determinant of a matrix SARRUS' RULE

Next, let us show you the way to evaluate a 3rd order determinant by Sarrus' rule.

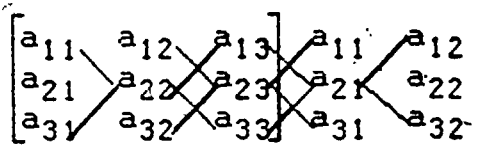
Copy the first two columns over again, and connect each of the three first-row elements of A with the two numbers located "southeast" of it by solid lines.

Similarily, connect each of the three third row elements by THICK solid lines with the two numbers located "northeast" of it.

Note that solid lines produce epsilon value +1, and the thick lines produce -1.

Thus, the value of determinant A is

$$+ a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}.$$



Press -NEXT-

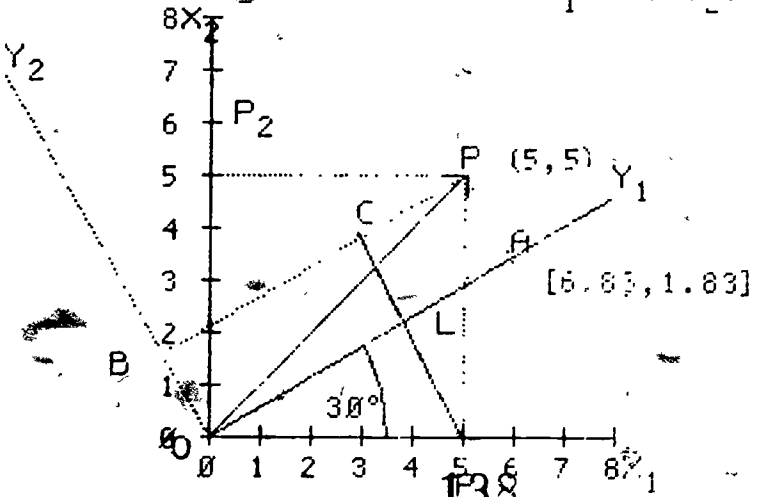
We just obtained the relations

$$OA = (\cos 30°) \times 5 + (\sin 30°) \times 5$$
$$OB = (-\sin 30°) \times 5 + (\cos 30°) \times 5$$

Substituting cos30°, sin30° by their values,

$$OA = .866 \times 5 + .5 \times 5 = 6.83$$
$$OB = -.5 \times 5 + .866 \times 5 = 1.83$$

Thus, OP is represented by [6.83, 1.83] using the new axes Y_1 and Y_2.



Press NEXT

APPENDIX B

PRETEST FOR MATRIX ALGEBRA

Those who have little or no background in matrix algebra may be unable to answer many of the items below. You may skip by pressing the NEXT key without answering.

You may then come back to the test after taking one or more lessons.

This test will provide you with some feedback so that you may choose only the lessons you have to learn from five lessons in the index.

to start.....press -NEXT-

1) Choose the right answer.

$$\begin{bmatrix} 3 & 7 \\ -5 & 7 \end{bmatrix} + \begin{bmatrix} -1 & 8 \\ 3 & 9 \end{bmatrix} = ?$$

a) $\begin{bmatrix} 4 & -1 \\ -8 & -2 \end{bmatrix}$ b) $\begin{bmatrix} 2 & 15 \\ -2 & 16 \end{bmatrix}$ c) $\begin{bmatrix} -4 & 15 \\ -8 & 16 \end{bmatrix}$ d) $\begin{bmatrix} 10 & 7 \\ -2 & 12 \end{bmatrix}$

2) Choose the right answer.

$$\begin{bmatrix} -1 & 2 \\ 6 & 9 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ -1 & -6 \end{bmatrix} = ?$$

a) $\begin{bmatrix} -2 & 2 \\ 7 & 15 \end{bmatrix}$ b) $\begin{bmatrix} 0 & 2 \\ 5 & 3 \end{bmatrix}$ c) $\begin{bmatrix} 0 & 5 \\ 2 & 3 \end{bmatrix}$ d) $\begin{bmatrix} 0 & 2 \\ 3 & 3 \end{bmatrix}$

3) Choose the right answer.

$$\begin{bmatrix} 8 & -1 \\ 0 & -7 \end{bmatrix} - \begin{bmatrix} -1 & 3 \\ -5 & -1 \end{bmatrix} = ?$$

a) $\begin{bmatrix} 9 & -4 \\ 5 & -6 \end{bmatrix}$ b) $\begin{bmatrix} 7 & 2 \\ -5 & -8 \end{bmatrix}$ c) $\begin{bmatrix} -4 & 15 \\ -8 & 16 \end{bmatrix}$ d) $\begin{bmatrix} 10 & 7 \\ -2 & 12 \end{bmatrix}$

4) Choose the right answer.

$$\begin{bmatrix} -1 & -1 \\ 2 & 6 \end{bmatrix} - \begin{bmatrix} 6 & 10 \\ 9 & -1 \end{bmatrix} = ?$$

a) $\begin{bmatrix} -7 & -11 \\ -7 & 7 \end{bmatrix}$ b) $\begin{bmatrix} 5 & 9 \\ 11 & 5 \end{bmatrix}$ c) $\begin{bmatrix} -7 & -7 \\ -11 & 7 \end{bmatrix}$ d) $\begin{bmatrix} -7 & -1 \\ 7 & 7 \end{bmatrix}$

5) Choose the right answer.

$$\begin{bmatrix} 8 & -1 \\ 0 & -7 \end{bmatrix} \times 10 = ?$$

a) $\begin{bmatrix} 80 & -1 \\ 0 & -7 \end{bmatrix}$ b) $\begin{bmatrix} 80 & -1 \\ 0 & -70 \end{bmatrix}$ c) $\begin{bmatrix} 80 & -10 \\ 0 & -70 \end{bmatrix}$ d) $\begin{bmatrix} 80 & -1 \\ 0 & -7 \end{bmatrix}$

6) Choose the right answer.

$$\begin{bmatrix} 7 & 2 \\ -8 & -2 \end{bmatrix} \times 5 = ?$$

a) $\begin{bmatrix} 35 & 2 \\ -8 & -10 \end{bmatrix}$ b) $\begin{bmatrix} 35 & 10 \\ -40 & -10 \end{bmatrix}$ c) $\begin{bmatrix} 35 & -40 \\ 10 & -10 \end{bmatrix}$ d) $\begin{bmatrix} 35 & 2 \\ -10 & -10 \end{bmatrix}$

7) Choose the right answer.

$$\begin{bmatrix} 2 & 4 \\ 9 & -1 \end{bmatrix} = ?$$

a) $\begin{bmatrix} -1 & 4 \\ 9 & 2 \end{bmatrix}$ b) $\begin{bmatrix} 1 & 4 \\ 9 & -2 \end{bmatrix}$ c) $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$ d) $\begin{bmatrix} 2 & 9 \\ 4 & -1 \end{bmatrix}$

8) Choose the right answer.

$$\left[\begin{bmatrix} -5 & -2 \\ 6 & 6 \end{bmatrix} + \begin{bmatrix} 2 & 5 \\ -4 & -4 \end{bmatrix}\right]$$

a) $\begin{bmatrix} -3 & 3 \\ 2 & 2 \end{bmatrix}$ b) $\begin{bmatrix} 2 & 2 \\ 3 & -1 \end{bmatrix}$ c) $\begin{bmatrix} -3 & 2 \\ 3 & 2 \end{bmatrix}$ d) $\begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$

9) Choose the right answer.

$$\left[-9 \times \begin{bmatrix} 7 & -5 \\ 8 & 2 \end{bmatrix}\right] = ?$$

a) $\begin{bmatrix} -63 & 45 \\ -72 & -18 \end{bmatrix}$ b) $\begin{bmatrix} 63 & 45 \\ -72 & 18 \end{bmatrix}$ c) $\begin{bmatrix} -1.3 & 1.8 \\ -1.1 & -4.5 \end{bmatrix}$ d) $\begin{bmatrix} -63 & -72 \\ 45 & -18 \end{bmatrix}$

10) Choose the right answer.

$$\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

a) $\begin{bmatrix} 1/(-1) & 0 \\ 0 & 1/(-1) \end{bmatrix}$ b) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ c) $\begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$ d) $\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$

11) What is the order of the product of

$$\begin{bmatrix} -4 & 5 \\ 7 & -1 \end{bmatrix} \times \begin{bmatrix} 8 & -3 \\ -8 & -1 \\ 4 & 18 \end{bmatrix} = ?$$

a) 2 x 2 matrix
b) 3 x 2 matrix
c) 3 x 3 matrix
d) 2 x 3 matrix
e) not computable

12) How many two-factor products involving A,B and their transpose are computable? (e.g., AB', BA' and B^2)

$$A = \begin{bmatrix} 5 & 4 \\ -3 & 4 \end{bmatrix} \quad B = \begin{bmatrix} 5 & 10 \\ -1 & 1 \\ 19 & 5 \end{bmatrix}$$

a) none b) 1 c) 2 d) 6 e) more than 6

13) Suppose a matrix A is 2 x 2 symetric matrix, choose the letter whose statement is not true.

a) $A = A'$
b) A is a square matrix
c) $AB = BA$ for any 2 x 2 matrix B
d) If the inverse of A, A^{-1} exists then $A^{-1} = (A')^{-1}$

14) C = AB where A is p x q, B is s x t. Which of the following statements is not necessarily true?

a) The order of C is p x t b) p = s c) s = q

15) Choose the correct answer.

$$\begin{bmatrix} 3 & -3 \\ 1 & 0 \end{bmatrix} \times \begin{bmatrix} 3 & 3 & 1 \\ 0 & -2 & 3 \end{bmatrix}$$

a) $\begin{bmatrix} 9 & 15 & -6 \\ 3 & 3 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 9 & 15 & -6 \\ 3 & 4 & 5 \end{bmatrix}$

c) $\begin{bmatrix} 9 & 9 & 3 \\ 0 & 6 & -9 \end{bmatrix}$ d) $\begin{bmatrix} 12 & 12 & 4 \\ 0 & 4 & -6 \end{bmatrix}$

16) Choose the right answer.

$$\begin{bmatrix} -1 & 0 & 6 \\ 2 & -3 & 5 \end{bmatrix} \begin{bmatrix} -2 & 0 \\ 4 & -1 \\ 0 & 1 \end{bmatrix} = ?$$

a) $\begin{bmatrix} -2 & 6 \\ 12 & -3 \end{bmatrix}$ b) $\begin{bmatrix} -2 & 6 \\ 4 & -1 \end{bmatrix}$

c) $\begin{bmatrix} 2 & 6 \\ -16 & 8 \end{bmatrix}$ d) $\begin{bmatrix} 4 & 6 \\ -16 & 5 \end{bmatrix}$

17) Two 3-dimensional vectors $u' = (u_1, u_2, u_3)$, $v' = (v_1, v_2, v_3)$, and a 3 x 3 matrix A are given. Choose the wrong statement.

a) u'v is a number
b) uv' is a number
c) u v is a number
d) vu' is a matrix
e) v'A'u is a number

18) If $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, its determinant $|A|$ is equal to

a) - ab + bc
b) a + d
c) ad - bc
d) ac - bd
e) a + b + c + d

19) The cofactors of the elements of the first row of $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ are

a) d, -c b) d, c c) d, -b d) b, c e) b, -c

20) The cofactors of the elements of the 1st column of $\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$ are

a) $\begin{vmatrix} e & f \\ h & i \end{vmatrix}$, $-\begin{vmatrix} b & c \\ h & i \end{vmatrix}$, $\begin{vmatrix} b & c \\ e & f \end{vmatrix}$

b) $-\begin{vmatrix} e & f \\ h & i \end{vmatrix}$, $\begin{vmatrix} b & c \\ h & i \end{vmatrix}$, $-\begin{vmatrix} b & c \\ e & f \end{vmatrix}$

c) $\begin{vmatrix} d & e \\ g & h \end{vmatrix}$, $-\begin{vmatrix} d & f \\ g & i \end{vmatrix}$, $\begin{vmatrix} d & e \\ g & h \end{vmatrix}$

d) $\begin{vmatrix} a & b \\ d & e \end{vmatrix}$, $-\begin{vmatrix} a & c \\ g & i \end{vmatrix}$, $\begin{vmatrix} a & b \\ d & e \end{vmatrix}$

e) a, -b, c

21) For a given matrix $A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$

choose the correct statement.

a) $d \begin{vmatrix} e & f \\ h & i \end{vmatrix} - e \begin{vmatrix} d & f \\ g & i \end{vmatrix} + f \begin{vmatrix} d & e \\ g & h \end{vmatrix} = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}$

b) $a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} = 0$

c) $a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix}$

d) $a \begin{vmatrix} e & f \\ h & i \end{vmatrix} - b \begin{vmatrix} d & f \\ g & i \end{vmatrix} + c \begin{vmatrix} d & e \\ g & h \end{vmatrix} = a + e + i$

22) If $A + B = A + C$ then $B = C$

a) true b) false

23) If $AB = AC$ then $B = C$

a) true b) false

24) If $AB = 0$ then necessarily $A = 0$ or $B = 0$

a) true b) false

25) $AB = BA$ for any matrices A and B

a) true b) false

26) $A (B + C) = AB + AC$

a) true b) false

27) $(A + B)' = B' + A'$

a) true b) false

28) $(AB)' = A'B'$

a) true b) false

29) If A = A' then AA' = I

a) true b) false

30) If P is invertible and $B = P^{-1} A P$, then the determinants of B and A are equal.

a) true b) false

31) Let A, B, C, D be n x n matrices, then the determinant of 2n x 2n matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ is the determinant of matrix AD - BC.

a) true b) false c) I don't know

32) Choose the right answer. The adjoint matrix of $A = \begin{bmatrix} -9 & -.33 \\ .5 & 2.1 \end{bmatrix}$ is

a) $\begin{bmatrix} -9.00 & 0.50 \\ -0.33 & 2.10 \end{bmatrix}$ b) $\begin{bmatrix} 2.10 & -0.33 \\ 0.50 & -9.00 \end{bmatrix}$

c) $\begin{bmatrix} 2.10 & 0.33 \\ -0.50 & -9.00 \end{bmatrix}$ d) $\begin{bmatrix} 2.10 & -0.50 \\ 0.33 & -9.00 \end{bmatrix}$

33) Choose the inverse of the triangular matrix. $\begin{bmatrix} 3 & 0 & 0 \\ -2 & 5 & 0 \\ 1 & -6 & -2 \end{bmatrix}$

a) $\begin{bmatrix} 3 & 0 & 0 \\ -2 & 5 & 0 \\ 1 & -6 & -2 \end{bmatrix}$ b) $\begin{bmatrix} 1/3 & 0 & 0 \\ -2 & 1/5 & 0 \\ 1 & -6 & -1/2 \end{bmatrix}$

c) $\begin{bmatrix} 1/3 & -2 & 1 \\ 0 & 1/5 & -2 \\ 0 & 0 & 1 \end{bmatrix}$ d) $\begin{bmatrix} 1/3 & 0 & 0 \\ 4/30 & 1/5 & 0 \\ -7/30 & -3/5 & -1/2 \end{bmatrix}$

34) Which one of the following has orthogonal row vectors?

a) $\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ b) $\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ c) $\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}$ d) $\begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$

35) Which of the following transformation matrices is not orthogonal?

a) $\begin{bmatrix} 1/\sqrt{5} & 2/\sqrt{5} \\ 2/\sqrt{5} & -1/\sqrt{5} \end{bmatrix}$ b) $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$ c) $\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$

d) $\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$ e) $\begin{bmatrix} 3/\sqrt{10} & 1/\sqrt{10} \\ 1/\sqrt{10} & 3/\sqrt{10} \end{bmatrix}$

36) The product of two orthogonal transformations is an orthogonal transformation matrix.

a) true b) false

37) The row vectors contained in an orthogonal transformation matrix are mutually orthogonal but are not necessary of unit length.

a) true b) false

38) The column vectors contained in an orthogonal transformation matrix are not mutually orthogonal when the row vectors are mutually orthogonal.

a) true b) false

39) Any rigid rotation is an orthogonal transformation matrix.

a) true b) false

40) An orthogonal transformation of axes will not change the length of vectors in the space.

a) true b) false

41) Suppose matrix $\Sigma = \begin{bmatrix} 5.3 & .5 \\ .5 & 10.1 \end{bmatrix}$ is a variance-covariance matrix.

Choose the wrong statement.

a) The characteristic equation of Σ is $|\Sigma - \lambda I| = 0$

b) The characteristic equation of Σ is $\lambda^2 - 15.4\lambda + |\Sigma| = 0$

c) Σ is always transformed by some matrix into the diagonal form.

d) The roots of the characteristic equation might be complex variables.

For items 42-46, assume the following:

Suppose Σ is a n x n variance-covariance matrix, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ are its characteristic roots (or eigenvalues), $v_1, v_2, \ldots v_n$ are the characteristic vectors (or eigen vectors) of Σ associated with $\lambda_1, \lambda_2, \ldots \lambda_n$ respectively.

42) If the rank of Σ is n, then the eigenvectors $v_1, v_2, \ldots, v_n$ are linearly independent.

a) true b) false

43) If $\lambda_1, \lambda_2 \ldots \lambda_n$ are of distinct values then λ_1 is the largest variance of any linear combination of x_i with fixed norm of the coefficient vector.

a) true b) false

44) Some of the eigenvalues may be negative.

a) true b) false

45) v_i, v_j are not mutually orthogonal.

a) true b) false

46) Choose the correct answer.

a) The constant term of the characteristic equation of Σ is the trace of Σ.
b) The constant term of the characteristic equation of Σ is the determinant of Σ.
c) The constant term of the characteristic equation of Σ is the determinant of adjoint matrix Σ.
d) None of the above is correct.

47) How are the eigenvalues and eigenvectors of Σ^{-1} related to those of Σ?

a) The eigenvalues of Σ^{-1} are the same as those of Σ, but the eigenvectors are inverted.
b) The eigenvalues of Σ^{-1} are reciprocals of those of Σ, and the eigenvectors are inverted.
c) The eigenvalues of Σ^{-1} are reciprocals of those of Σ, but the eigenvectors are unchanged.
d) Both the eigenvalues and eigenvectors of Σ^{-1} are respectively the same as those of Σ.

48) If a 2 x 2 matrix A has eigenvalues λ_1, λ_2, then the eigenvalues of kA (where k is a scalar) are

a) $k^2\lambda_1$, $k^2\lambda_2$

b) $k^{1/2}\lambda_1$, $k^{1/2}\lambda_2$

c) λ_1, λ_2

d) $k\lambda_1$, $k\lambda_2$

e) $k\lambda_1$, λ_2

You have completed the test.

Press -BACK- if you wish to review your work and make changes.

Press -NEXT- to review the test and to see the correct answers.

Appendix C

Description of Contents in the Lessons of Chanute

lesson	average time	Content
103	33.27	Principles of Gas Engine
104a	34.28	Identification of Parts and Purpose of Gasoline Engine Compressor
104b	missed	
105	44.74	Cooling System
201a	12.55	Air and Exhaust System
201b	42.31	
202a	189.63	Fundamentals of Electricity
202b	14.24	Batteries
203a		
203b	missed	Electrical Schematics
203c		
204	100.20	Starters
205a	136.51	Cranking Motors, DC Charging System
205b	41.20	AC Charging System
206a	50.22	
206b	21.43	Battery Ignition
206c	43.69	
207	37.77	Emission Control
301	32.40	Diesel Engines
303	14.04	Lighting System
304	12.81	Warning System
305	22.56	Clutches
307	72.67	Basic Hydraulics
308	46.60	Fluid Couplings/Torque Converters
401	20.84	V-Joints/Propeller Shafts
402	91.09	Differentials
403	13.35	Transfer Case/PTO
404	12.60	Suspension System
405a	31.17	Hydraulic and Mechanical Brakes
405b	52.96	Air Brakes
405c	13.64	Power Assisted Brakes

Appendix D

Description of PLATO Programs and their Programmers

Lesson Name	Programmer and Description
matx4	Jim Kraatz Test items were developed by one of the authors but the test frame and data collection scheme was developed by Jim Kraatz of CERL. Up to 50 items can be handled and item scores, response time for each item, and selected option of multiple choice are collected. The traditional item analysis, such as means, discriminating powers of each item are given.
edittest	Robert Baillie, Jim Kraatz Routine for editing data fromthe "matx4" test driver.
storetest	Robert Baillie Transformation routines. This program prepares the data from "matx4" test driver for various analysis such as "datam," "wb2," and "Kolmo."
gram	Robert Baillie Orthogonalize up to 10 vectors by Gram-Schmidt method and estimates an individual student's gain scores. Eight vectors (variables) besides the pre-test and post-test can be used to step up the accuracy of the gain scores.
subr	Jerry Dyer and Robert Baillie Calculates various probability functions. They can be used as a statistical table by condensing this lesson, but they are mainly used as subroutines in user's programs. This program contains F and F^{-1} distributions, X^2 and inverse X^2, normal and inverse normal disributions, t, distribution, binomial, beta, incomplete beta, and two parameter gamma disributions.
matsubr	Jerry Dyer, Robert Baillie, and Kay Tatsuoka Calculates the inverse, eigenvalues and eigenvectors, and determinant of a 20 x 20 matrix.
Kolmonorm	Robert Baillie and Jerry Dyer Kolmogorov-Smirnov test of a sample and a given theoretical distribution function, such as Weibull, Gamma, normal distributions. Uses "statedit" to input data.
cutoff	Tamar Weaver Evaluates the optimum cutoff scores of a criterion-referenced test and calculates the estimation of false positive, negative, failure rate, success rate based on a user's specified true mastery level and observed cutoff score. Classify an individual's score into one of four status groups: pure pass, fail, false positive, or false negative.

multreg, multrega, multreg2, multreg3, and linkage program to statedit, formatf	Kumi Tatsuoka, Robert Baillie, Tamar Weaver Input raw data and matrix into temporary storage, calculate a correlation matrix up to 20x20, partial correlations and stepwise multiple regression. The data stored in a dataset via Felty's "statedit" is acceptable.
lintest	Kay Tatsuoka Tests linearity of the data.
manova	Kay Tatsuoka, Robert Baillie Multivariate analysis of variance
sscp	Robert Baillie Discriminant analysis for one factor, several groups and variables, using a dataset with "statedit" data format.
sscp2	Robert Baillie Discriminant analysis with temporary storage.
factdisc	Kay Tatsuoka Factorial discriminant analysis.
ccor	Robert Baillie Canonical correlation analysis -- takes data stored in "statedit" format, need a dataset.
ccor2	Robert Baillie Canonical correlation analysis using temporary storage.
varimax	Kay Tatsuoka, Robert Baillie Do principal component analysis and rotate a factor matrix by Varimax rotation.
area package	Tamar Weaver, Al Avner, Kumi Tatsuoka Collect the area data specified by an author in his/her lesson.
formatk	Tamar Weaver Transforms area data in a "statedit" format dataset.
kst1	Kay Tatsuoka Augments data from several different datasets which are stored in the "statedit" data format.
chitest	Kay Tatsuoka Do a simple factorial analysis of variance and X^2 goodness of fit test.
mdc1	Mark Bradley Edit and simple analysis of the quizzes and tests in the matrix algebra lessons. These data were not used in the report.

Appendix E

Tables of p-values and the Weibull Parameters

Table E1

Kolmogorov-Smirnov Tests for Matrix Algebra Pretest Items for OK subgroup

item	p	z	N	item	p	z	N
1)	Ø.ØØ616	1.3191	9Ø	25)	Ø.ØØ6Ø7	1.3219	62
2)	Ø.ØØ211	1.5ØØ2	96	26)	Ø.ØØØØ1	2.262Ø	81
3)	Ø.6417	Ø.7414	78	27)	Ø.1Ø97	1.2Ø48	49
4)	Ø.5115	Ø.82Ø3	79	28)	Ø.5Ø71	Ø.8231	36
5)	Ø.4456	Ø.8631	94	29)	Ø.9999	Ø.3234	26
6)	Ø.1558	1.1296	96	3Ø)	Ø.6781	Ø.7198	24
7)	Ø.2213	1.Ø488	61	31)	Ø.3388	Ø.94Ø9	31
8)	Ø.6Ø89	Ø.76Ø9	54	32)	Ø.7Ø88	Ø.7Ø15	33
9)	Ø.3Ø63	Ø.9677	7Ø	33)	1.ØØØØ	Ø.2735	29
1Ø)	Ø.9228	Ø.55ØØ	65	34)	Ø.6Ø92	Ø.76Ø7	27
11)	Ø.5Ø59	Ø.8238	34	35)	Ø.17Ø6	1.1Ø93	57
12)	Ø.9215	Ø.5513	3Ø	36)	Ø.3154	Ø.96ØØ	37
13)	Ø.9137	Ø.5588	43	37)	Ø.9646	Ø.4989	9
14)	Ø.7459	Ø.679Ø	42	38)	Ø.6187	Ø.7551	23
15)	Ø.1285	1.1716	49	39)	Ø.7291	Ø.6892	16
16)	Ø.8567	Ø.6Ø57	6Ø	4Ø)	Ø.1785	1.Ø989	28
17)	Ø.8832	Ø.5852	21	41)	Ø.9731	Ø.4844	21
18)	Ø.9587	Ø.5Ø78	61	42)	Ø.5664	Ø.7864	2Ø
19)	Ø.8643	Ø.6ØØØ	26	43)	Ø.7422	Ø.6812	18
2Ø)	Ø.5751	Ø.7811	46	44)	Ø.94Ø7	Ø.53Ø9	16
21)	Ø.965Ø	Ø.4984	42	45)	Ø.8249	Ø.6282	12
22)	Ø.ØØ65	1.6928	77	46)	1.ØØØØ	Ø.32ØØ	16
23)	Ø.7124	Ø.6993	16	47)	Ø.2973	Ø.9754	4Ø
24)	Ø.9942	Ø.4218	27	48)	Ø.8687	Ø.5967	32

Pretest for all subjects after 1976 Fall semester; 'goodness of fit' testing for Weibull distributions

Table E2

The Three Weibull Parameters for Matrix Algebra Test Items

items	t_0	m.c.	c	μ_0
1.	10.59	0.96	1.15	44.51
2.	2.53	0.96	1.92	28.12
3.	6.02	0.99	1.24	23.34
4.	0.00	0.98	2.19	32.35
5.	7.54	0.99	1.25	13.70
6.	5.51	0.98	1.35	12.67
7.	7.52	0.99	1.15	33.92
8.	0.00	0.98	1.56	57.62
9.	4.00	0.99	1.71	40.68
10.	7.27	1.00	1.45	22.00
11.	4.49	0.98	1.33	41.95
12.	8.24	0.99	0.97	57.67
13.	5.41	0.99	1.18	57.21
14.	9.29	0.99	1.34	43.46
15.	0.00	0.97	1.08	112.06
16.	0.45	0.99	1.26	57.26
17.	15.24	0.99	1.14	108.62
18.	2.77	1.00	1.63	21.67
19.	0.00	0.99	1.24	33.70
20.	1.37	0.99	1.00	49.98
21.	5.24	1.00	1.04	49.83
22.	0.61	0.94	1.90	23.27
23.	3.19	0.95	1.02	26.15
24.	4.60	1.00	1.14	13.91
25.	3.47	0.98	1.42	11.04
26.	1.96	0.93	0.87	13.70

Table E2 (con't)

The Three Weibull Parameters for Matrix Algebra Test Items

Items	t_0	m.c.	c	μ_0
27.	2.94	0.98	0.84	15.31
28.	3.94	0.98	0.62	9.89
29.	6.56	1.00	0.97	25.67
30.	5.72	0.97	0.85	28.69
31.	1.52	0.97	1.39	32.93
32.	2.50	0.99	0.91	20.99
33.	0.00	1.00	1.04	138.67
34.	5.51	0.98	0.98	30.55
35.	1.96	0.98	1.71	52.95
36.	3.89	0.98	0.91	13.15
37.	5.65	0.97	0.93	13.86
38.	6.74	1.00	1.46	17.04
39.	3.12	0.98	1.49	9.73
40.	3.77	0.98	1.10	9.76
41.	4.21	0.99	1.79	38.05
42.	3.46	0.93	1.26	31.78
43.	7.28	0.99	1.42	21.32
44.	3.84	0.98	1.35	15.98
45.	3.34	0.99	1.71	17.20
46.	1.70	0.99	1.31	30.64
47.	1.21	0.99	1.22	16.61
48.	5.17	0.99	1.19	15.63

*Pretest given after 76 Fall semester, OK subgroup

Table E3

Kolmogorov-Smirnov Tests for Matrix Algebra Pretest Items for NO subgroup

item	p	z	N	item	p	z	N
1)	Ø.9912	Ø.4364	1Ø	25)	Ø.2976	Ø.9751	38
2)	1.ØØØØ	Ø.2797	4	26)	Ø.9776	Ø.4754	19
3)	Ø.4Ø68	Ø.8899	22	27)	Ø.1566	1.1285	51
4)	Ø.6647	Ø.7277	21	28)	Ø.ØØ88	1.395Ø	64
5)	Ø.9882	Ø.4474	5	29)	Ø.ØØ17	1.5Ø35	74
6)	Ø.42Ø3	Ø.88Ø4	4	3Ø)	Ø.ØØ25	1.513Ø	76
7)	Ø.7219	Ø.6936	38	31)	Ø.9192	Ø.5535	69
8)	Ø.9458	Ø.5249	46	32)	Ø.73Ø5	Ø.6884	67
9)	Ø.9Ø2Ø	Ø.5694	3Ø	33)	Ø.6574	Ø.7321	71
1Ø)	Ø.96Ø8	Ø.5Ø48	34	34)	Ø.5696	Ø.7845	73
11)	Ø.1987	1.Ø743	66	35)	Ø.5377	Ø.8Ø4Ø	43
12)	Ø.8269	Ø.6268	7Ø	36)	Ø.ØØ87	1.4Ø43	63
13)	Ø.8959	Ø.5747	57	37)	Ø.79Ø7	Ø.65Ø9	9Ø
14)	Ø.96Ø3	Ø.5Ø56	58	38)	Ø.2Ø83	1.Ø632	76
15)	Ø.6791	Ø.7192	51	39)	Ø.3499	Ø.9321	83
16)	Ø.9418	Ø.5296	4Ø	4Ø)	Ø.4919	Ø.8328	71
17)	Ø.8424	Ø.616Ø	79	41)	Ø.4429	Ø.8649	78
18)	Ø.8Ø33	Ø.6426	39	42)	Ø.7938	Ø.6488	79
19)	Ø.7451	Ø.6795	74	43)	Ø.ØØ25	1.28ØØ	81
2Ø)	Ø.2778	Ø.9928	54	44)	Ø.ØØØ8	1.6311	83
21)	Ø.Ø1Ø8	1.6162	58	45)	Ø.ØØØØ	2.4586	87
22)	Ø.9Ø68	Ø.5652	23	46)	Ø.3486	Ø.9332	83
23)	Ø.ØØ97	1.6327	84	47)	Ø.1819	1.Ø947	58
24)	Ø.Ø291	1.4545	73	48)	Ø.2734	Ø.9968	66

Pretest for all subjects after 1976 Fall semester; 'goodness of fit' testing for Weibull distributions

Table E4

The Three Weibull Parameters for Matrix Algebra Test Items

items	t_0	m.c.	c	μ_0
1.	0.00	0.99	2.08	52.64
2.	11.49	1.00	0.49	30.09
3.	6.50	0.96	1.05	19.29
4.	7.65	0.00	1.02	14.47
5.	5.64	1.00	2.10	17.45
6.	0.00	0.90	0.85	10.79
7.	6.24	0.99	1.44	37.43
8.	0.00	0.99	1.33	40.45
9.	0.28	0.99	1.14	27.68
10.	2.54	1.00	1.11	20.96
11.	0.95	0.97	1.45	30.27
12.	3.21	1.00	1.13	58.24
13.	0.00	0.99	1.60	47.01
14.	2.06	1.00	1.24	45.35
15.	5.74	0.90	0.82	64.92
16.	0.53	0.99	0.93	36.65
17.	0.55	0.99	1.13	41.25
18.	1.07	0.99	1.34	24.76
19.	2.28	1.00	1.38	26.12
20.	3.89	0.97	0.89	22.03
21.	0.40	0.96	1.22	34.10
22.	2.29	0.99	1.17	17.08
23.	1.48	0.97	1.56	16.30
24.	3.45	0.98	1.59	8.62
25.	2.62	0.90	1.08	11.00
26.	0.42	0.99	1.59	14.79

Table E4 (con't)

The Three Weibull Parameters for Matrix Algebra Test Items

items	t_0	m.c.	c	μ_0
27.	0.37	0.98	1.64	11.52
28.	1.98	0.98	0.80	7.12
29.	1.88	0.98	1.12	10.86
30.	1.85	0.99	1.13	13.97
31.	2.85	1.00	0.95	13.14
32.	4.61	1.00	1.10	35.26
33.	1.75	0.99	0.98	44.33
34.	0.00	0.99	2.10	24.06
35.	2.39	0.99	1.13	29.62
36.	1.83	0.97	1.28	8.41
37.	2.76	1.00	1.16	11.64
38.	2.91	0.99	0.81	8.30
39.	1.78	1.00	1.31	6.59
40.	1.80	0.99	1.11	6.63
41.	4.86	0.99	0.95	28.31
42.	4.76	1.00	0.74	10.66
43.	3.97	0.99	0.95	7.00
44.	2.97	0.97	0.94	7.19
45.	1.94	0.94	1.03	9.64
46.	1.64	0.99	1.34	16.74
47.	5.85	0.98	1.06	14.48
48.	1.49	0.99	1.20	17.99

*Pretest, given after 76 Fall semester, NO subgroup

Table E5

Kolmogorov-Smirnov Tests for Matrix Algebra Test Items for OK subgroup

item	p	z	N	item	p	z	N
1)	Ø.ØØØØ	2.7994	9Ø	25)	Ø.Ø481	1.3651	62
2)	Ø.ØØØØ	1.9841	96	26)	Ø.ØØØØ	4.7953	81
3)	Ø.1823	1.Ø942	78	27)	Ø.ØØ53	1.7213	49
4)	Ø.9139	Ø.5586	79	28)	Ø.ØØØ4	2.Ø688	36
5)	Ø.1Ø76	1.2Ø88	94	29)	Ø.8979	Ø.573Ø	26
6)	Ø.ØØ42	1.7557	96	3Ø)	Ø.ØØ95	1.6362	24
7)	Ø.1Ø12	1.2213	61	31)	Ø.2322	1.Ø373	31
8)	Ø.6397	Ø.7426	54	32)	Ø.6328	Ø.7467	33
9)	Ø.68Ø9	Ø.7181	7Ø	33)	1.ØØØØ	Ø.2884	29
1Ø)	Ø.8Ø39	Ø.6423	65	34)	Ø.Ø9ØØ	1.2452	27
11)	Ø.3451	Ø.9359	34	35)	Ø.2346	1.Ø348	57
12)	Ø.2275	1.Ø422	3Ø	36)	Ø.ØØ62	1.6985	37
13)	Ø.9137	Ø.5588	43	37)	Ø.869Ø	Ø.5964	9
14)	Ø.6949	Ø.7Ø98	42	38)	Ø.56Ø5	Ø.79ØØ	23
15)	Ø.23Ø6	1.Ø389	49	39)	Ø.76Ø6	Ø.6699	16
16)	Ø.916Ø	Ø.5566	6Ø	4Ø)	Ø.Ø329	1.4332	28
17)	Ø.78Ø4	Ø.6575	21	41)	Ø.8993	Ø.5718	21
18)	Ø.9689	Ø.492Ø	61	42)	Ø.Ø178	1.5363	2Ø
19)	Ø.5555	Ø.7931	26	43)	Ø.76Ø7	Ø.6698	18
2Ø)	Ø.551Ø	Ø.7958	46	44)	Ø.9579	Ø.5Ø9Ø	16
21)	Ø.8376	Ø.6194	42	45)	Ø.8395	Ø.6181	12
22)	Ø.ØØ37	1.7754	77	46)	Ø.9985	Ø.3845	16
23)	Ø.Ø745	1.2826	16	47)	Ø.365Ø	Ø.92Ø5	4Ø
24)	Ø.9468	Ø.5236	27	48)	Ø.7982	Ø.646Ø	32

Pretest for all subjects after 1976 Fall semester: 'goodness of fit' testing for Gamma distributions

Table E6

Kolmogorov-Smirnov Tests for Matrix Algebra Test Items for NO subgroup

item	p	z	N	item	p	z	N
1)	Ø.9996	Ø.3540	10	25)	Ø.0805	1.2675	38
2)	Ø.9623	Ø.5026	4	26)	Ø.8151	Ø.6348	19
3)	Ø.1775	1.1003	22	27)	Ø.3833	Ø.9069	51
4)	Ø.0748	1.2817	21	28)	Ø.0000	2.9113	64
5)	Ø.9592	Ø.5071	5	29)	Ø.0028	1.8111	74
6)	Ø.3186	Ø.9573	4	30)	Ø.0238	1.4883	76
7)	Ø.7544	Ø.6737	38	31)	Ø.1990	1.0739	69
8)	Ø.9885	Ø.4463	46	32)	Ø.9090	Ø.5631	67
9)	Ø.7421	Ø.6813	30	33)	Ø.0001	2.2033	71
10)	Ø.9417	Ø.5298	34	34)	Ø.8965	Ø.5742	73
11)	Ø.0010	1.9456	66	35)	Ø.5527	Ø.7948	43
12)	Ø.8839	Ø.5847	70	36)	Ø.0011	1.9415	63
13)	Ø.8749	Ø.5918	57	37)	Ø.3386	Ø.9411	90
14)	Ø.9500	Ø.5196	58	38)	Ø.0000	2.5784	76
15)	Ø.0000	2.6823	51	39)	Ø.2728	Ø.9974	83
16)	Ø.9646	Ø.4990	40	40)	Ø.0279	1.4616	71
17)	Ø.8622	Ø.6016	79	41)	Ø.2250	1.0449	78
18)	Ø.9316	Ø.5410	39	42)	Ø.0026	1.8225	79
19)	Ø.7355	Ø.6853	74	43)	Ø.0003	2.1121	81
20)	Ø.0027	1.8167	54	44)	Ø.0000	2.9649	83
21)	Ø.0966	1.2309	58	45)	Ø.0000	3.6842	87
22)	Ø.9124	Ø.5600	23	46)	Ø.0683	1.2993	83
23)	Ø.0645	1.3103	84	47)	Ø.0075	1.6713	58
24)	Ø.0206	1.5126	73	48)	Ø.2291	1.0405	66

pretest after Fall 76 : fitting Gamma

Table E7

Kolmogorov-Smirnov Tests for Matrix Algebra Pretest Items ; Matched Sample

NO Group				OK Group			
item	p	z	N	item	p	z	N
1)	1.0000	0.3196	5	1)	0.6927	0.7111	51
2)	0.2700	1.0000	1	2)	0.2707	0.9994	55
3)	0.1690	1.1114	13	3)	0.8910	0.5789	43
4)	0.9456	0.5251	12	4)	0.9976	0.3962	44
5)	0.8928	0.5774	3	5)	0.7999	0.6449	53
6)	0.9996	0.3536	2	6)	0.1503	1.1375	54
7)	0.9928	0.4290	30	7)	0.8781	0.5894	26
8)	0.9922	0.4320	31	8)	0.7253	0.6916	25
9)	0.8941	0.5762	21	9)	0.8755	0.5914	35
10)	0.9250	0.5479	24	10)	0.8941	0.5762	32
11)	0.6932	0.7108	34	11)	0.9113	0.5611	22
12)	0.9954	0.4145	40	12)	0.7020	0.7055	16
13)	0.9375	0.5345	35	13)	0.9185	0.5542	21
14)	0.7875	0.6529	34	14)	0.9518	0.5173	22
15)	0.7963	0.6472	35	15)	0.7017	0.7057	21
16)	0.9873	0.4503	29	16)	0.9962	0.4089	27
17)	0.9303	0.5424	49	17)	0.9171	0.5556	7
18)	0.9100	0.5623	26	18)	0.9941	0.4225	30
19)	0.6220	0.7531	25	19)	0.4319	0.8725	31
20)	0.9818	0.4656	11	20)	0.0389	1.4036	45
21)	0.2851	0.9861	31	21)	0.5322	0.8074	25
22)	0.1638	1.1184	44	22)	0.5341	0.8062	12
23)	0.1910	1.0834	45	23)	0.9999	0.3228	11

Pretest for matched group after 1976 Fall semester; 'goodness of fit' testing for Weibull distributions

Table E8

The Three Weibull Parameters for Matrix Algebra Test Items

items	t_0	m.c.	c	μ_0
1.	9.87	Ø.97	1.44	33.67
2.	2.93	Ø.99	2.42	24.1Ø
3.	5.33	Ø.99	1.39	2Ø.21
4.	6.3Ø	1.ØØ	1.Ø5	23.73
5.	8.77	Ø.99	1.17	13.13
6.	5.21	Ø.99	1.54	11.89
7.	9.Ø6	1.ØØ	1.11	33.17
8.	Ø.ØØ	Ø.98	1.45	6Ø.88
9.	9.85	1.ØØ	1.55	32.95
1Ø.	8.14	Ø.99	1.63	22.43
11.	4.83	Ø.98	1.12	37.92
12.	8.89	Ø.98	Ø.99	36.55
13.	5.42	Ø.99	1.Ø7	52.52
14.	1Ø.29	1.ØØ	Ø.95	36.24
15.	24.94	Ø.98	Ø.99	95.1Ø
16.	5.62	1.ØØ	1.Ø8	5Ø.69
17.	Ø.ØØ	Ø.97	1.25	132.76
24.	3.48	Ø.98	1.33	12.32
25.	3.32	Ø.99	1.54	1Ø.59
26.	4.94	Ø.96	Ø.54	6.74
27.	3.94	Ø.98	Ø.64	9.89
28.	3.99	Ø.98	Ø.5Ø	7.1
29.	7.79	1.ØØ	Ø.64	24.64

Posttest for matched group, Multpost for QK subgroup

Table E9

The Three Weibull Parameters for Matrix Algebra Test Items

items	$t_{\emptyset}$	m.c.	c	$\mu_{\circ}$
1.	21.Ø4	1.ØØ	1.12	29.62
3.	7.45	Ø.93	Ø.9Ø	21.42
4.	5.18	Ø.99	2.54	13.54
5.	Ø.ØØ	1.ØØ	3.97	21.49
6.	Ø.ØØ	1.ØØ	****	13.79
[redacted]	5.69	1.ØØ	1.54	32.93
8.	6.Ø1	Ø.99	1.15	36.98
9.	3.67	Ø.99	Ø.82	27.4Ø
10.	1.91	Ø.99	1.31	22.44
11.	10.29	Ø.99	1.13	3Ø.31
12.	4.23	1.ØØ	1.37	45.23
13.	9.Ø9	Ø.98	1.23	32.31
14.	3.41	Ø.99	1.21	47.5Ø
15.	3.46	Ø.99	1.13	54.65
16.	3.61	1.ØØ	Ø.84	47.39
17.	4.73	Ø.99	Ø.98	32.Ø2
24.	3.27	Ø.98	1.56	9.46
25.	2.69	Ø.98	1.11	1Ø.84
26.	1.12	1.ØØ	1.34	11.82
27.	2.74	Ø.99	1.Ø7	6.79
28.	2.92	Ø.99	Ø.65	5.16
29.	2.76	Ø.97	1.ØØ	8.Ø3

Multpost for NO subgroup, matched sample

Table E10

Kolmogorov-Smirnov Tests for Matrix Algebra Posttest Items ; Matched Sample

NO Group				OK Group			
item	p	z	N	item	p	z	N
1)	0.0000	0.0000	0	1)	0.6305	0.7480	56
2)	0.0000	0.0000	0	2)	0.7303	0.6885	56
3)	0.8323	0.6231	7	3)	0.9344	0.5380	49
4)	0.0000	0.0000	0	4)	0.5666	0.7863	56
5)	0.9996	0.3536	2	5)	0.6345	0.7456	54
6)	0.0000	0.0000	0	6)	0.1768	1.1012	56
7)	0.0000	0.0000	0	7)	0.8277	0.6263	55
8)	0.0000	0.0000	0	8)	0.7035	0.7046	55
9)	0.9981	0.3899	4	9)	0.9364	0.5357	52
10)	0.0000	0.0000	0	10)	0.1114	1.2016	56
11)	0.7364	0.6848	36	11)	0.9354	0.5368	28
12)	0.9577	0.5093	27	12)	0.9788	0.4727	29
13)	0.9446	0.5263	16	13)	0.7324	0.6872	40
14)	0.7286	0.6895	12	14)	0.7929	0.6494	44
15)	0.9455	0.5252	7	15)	0.8850	0.5838	49
16)	0.9240	0.5489	4	16)	0.2204	1.0498	52
17)	0.5312	0.8080	30	17)	0.6169	0.7561	26
18)	0.9267	0.5461	21	18)	0.7893	0.6518	35
19)	0.9826	0.4636	4	19)	0.3674	0.9188	52
20)	0.9984	0.3850	5	20)	0.1832	1.0930	51
21)	0.9295	0.5432	7	21)	0.5858	0.7747	49
22)	0.9887	0.4457	21	22)	0.8147	0.6351	35
23)	0.7709	0.6635	36	23)	0.9030	0.5686	28

Posttest for matched group after 1976 Fall semester; 'goodness of fit' testing for Weibull distributions

Table E11

The Three Weibull Parameters for Matrix Algebra Test Items

items	t_0	m.o.	c	μ_o
1.	11.57	0.98	1.04	24.29
2.	8.13	1.00	1.73	13.43
3.	6.47	1.00	1.29	17.81
4.	14.13	0.99	1.32	17.75
5.	6.75	0.99	1.10	10.07
6.	9.66	0.99	1.38	11.76
7.	6.70	0.97	0.90	12.51
8.	15.29	0.99	1.11	27.72
9.	7.23	0.99	1.47	29.19
10.	6.73	0.96	1.01	16.91
11.	11.34	0.99	1.28	46.88
12.	1.12	0.99	1.15	110.78
13.	10.49	0.98	0.95	54.62
14.	0.32	0.99	1.65	58.94
15.	19.06	0.99	1.20	95.00
16.	8.59	0.98	1.41	77.43
17.	26.91	0.98	0.87	111.12
18.	5.72	1.00	0.93	26.24
19.	2.71	0.99	1.40	8.63
20.	2.94	0.97	0.80	8.58
21.	2.80	0.99	0.69	16.50
22.	3.76	0.99	0.74	17.34
23.	5.80	0.98	1.10	27.01

Posttest for matched sample, multpost of OK subgroup

Table E12

The Three Weibull Parameters for Matrix Algebra Test Items

Items	t_0	m.c.	c	μ_0
3.	1.46	0.97	1.75	16.02
5.	0.00	1.00	1.41	18.57
9.	12.16	0.99	0.76	26.18
11.	9.56	1.00	0.99	34.18
12.	12.73	0.99	1.22	85.71
13.	28.64	0.90	1.19	42.74
14.	16.51	0.99	1.48	38.64
15.	0.00	0.96	0.75	140.45
16.	0.00	0.96	1.04	149.16
17.	0.00	0.98	1.31	91.28
18.	5.63	0.99	0.75	25.11
19.	2.93	0.98	0.88	10.31
20.	2.54	0.99	1.26	14.10
21.	4.90	0.97	0.53	14.54
22.	3.70	1.00	0.76	6.44
23.	4.85	0.99	0.79	14.52

Multpost for NO subgroup in matched sample

Table E13

Kolmogorov-Smirnov Tests for Matrix Algebra Posttest Items ; Matched Sample

	NO subgroup			OK subgroup		
item	p	z	N	p	z	N
1)	0.0000	0.0000	0	0.0000	2.4473	56
2)	0.0000	0.0000	0	0.6675	0.7261	56
3)	0.9192	0.5535	7	0.9312	0.5414	49
4)	0.0000	0.0000	0	0.4610	0.8528	56
5)	0.9788	0.4727	2	0.0817	1.2646	54
6)	0.0000	0.0000	0	0.1957	1.0779	56
7)	0.0000	0.0000	0	0.0231	1.4935	55
8)	0.0000	0.0000	0	0.1042	1.2153	55
9)	0.9847	0.4580	4	0.7550	0.6733	52
10)	0.0000	0.0000	0	0.0000	3.4737	56
11)	0.3983	0.8960	36	0.7752	0.6607	28
12)	0.9298	0.5429	27	0.9940	0.4229	29
13)	0.8413	0.6168	16	0.2159	1.0547	48
14)	0.6815	0.7177	12	0.9332	0.5392	44
15)	0.8296	0.6250	7	0.2027	1.0697	49
16)	0.8272	0.6266	4	0.1619	1.1211	52
17)	0.7366	0.6847	30	0.1102	1.2039	26
18)	0.2239	1.0460	21	0.9032	0.5684	35
19)	0.9307	0.5419	4	0.1693	1.1110	52
20)	0.9971	0.4004	5	0.0000	3.6924	51
21)	0.1835	1.0927	7	0.1411	1.1513	49
22)	0.1523	1.1345	21	0.1772	1.1007	35
23)	0.0133	1.5837	36	0.9472	0.5231	28

Psottest for matched group, after 1976 Fall semester; goodness of fit testing for Gamma distributions

TABLE E14

Kolmogorov-Smirnov Tests for Matrix Algebra Posttest Items

OK Group				NO Group			
item	p	z	N	item	p	z	N
1)	0.3300	0.9480	64	1)			0
2)	0.2646	1.0051	64	2)			0
3)	0.9378	0.5342	55	3)	.4834	.8382	9
4)	0.1693	1.1109	63	4)			0
5)	0.3757	0.9125	62	5)	.9996	.3536	2
6)	0.0751	1.2810	64	6)			0
7)	0.6907	0.7123	61	7)	.0050	1.7321	3
8)	0.3863	0.9047	62	8)	.0366	1.4142	2
9)	0.6921	0.7115	57	9)	.6504	.7362	7
10)	0.0237	1.4891	64	10)			0
11)	0.8667	0.5982	23	11)	.7463	.6787	41
12)	0.9929	0.4285	31	12)	.9350	.5373	33
13)	0.8903	0.5794	47	13)	.9093	.5628	17
14)	0.7517	0.6754	50	14)	.7109	.7002	14
15)	0.9255	0.5473	56	15)	.6895	.7130	8
16)	0.2001	1.0726	59	16)	.6686	.7255	5
17)	0.4410	0.8662	30	17)	.7767	.6598	34
18)	0.8377	0.6194	36	18)	.5818	.7771	28
19)	0.3289	0.9489	59	19)	.6408	.7419	5
20)	0.1469	1.1425	59	20)	.9984	.3850	5
21)	0.5989	0.7668	53	21)	.7015	.7058	11
22)	0.8014	0.6439	38	22)	.9605	.5052	26
23)	0.9030	0.5686	20	23)	.3914	.9010	44
24)	0.0000	5.1962	27				

Posttest for all subjects after 1976 Fall semester; 'goodness of fit' testing for Weibull distributions

Table E15

The Three Weibull Parameters for Matrix Algebra Test Items

Items	t_0	m.c.	c	μ_0
1.	11.59	0.97	1.06	25.25
2.	8.00	0.97	2.76	28.61
3.	6.24	1.00	1.24	17.06
4.	9.82	0.96	1.05	20.87
5.	6.76	0.99	1.00	10.08
6.	5.67	0.99	1.41	11.02
7.	6.68	0.99	0.94	13.01
8.	15.46	0.99	1.05	30.36
9.	7.33	0.99	1.46	30.48
10.	5.78	0.96	0.97	18.72
11.	9.61	0.99	1.26	45.78
12.	0.28	0.99	1.00	117.45
13.	10.11	0.99	0.95	56.49
14.	1.42	0.99	1.51	60.41
15.	19.28	0.99	1.21	93.14
16.	8.51	0.99	1.48	77.99
17.	3.03	0.97	1.18	143.08
18.	5.74	1.00	0.93	25.47
19.	2.76	0.99	1.34	9.00
20.	2.94	0.97	0.82	9.17
21.	3.82	0.98	0.91	15.03
22.	3.82	0.99	0.71	17.05
23.	5.00	0.99	1.10	27.01

Multiple of OK subgroup in all subjects

Table E16

The Three Weibull Parameters for Matrix Algebra Test Items

items	t_0	m.c.	c	μ_0
1.	1.00	0.00	0.00	0/0
2.	1.00	0.00	0.00	0/0
3.	1.46	0.97	1.75	16.02
5.	0.00	1.00	4.41	18.57
9.	12.16	0.99	0.76	26.18
11.	9.56	1.00	0.99	34.10
12.	12.73	0.99	1.22	35.71
13.	28.64	0.98	1.19	42.74
14.	10.51	0.99	1.48	38.64
15.	0.00	0.96	0.75	140.55
16.	0.00	0.96	1.64	139.16
17.	0.00	0.98	1.31	91.28
18.	5.63	0.99	0.75	20.11
19.	2.93	0.98	0.88	10.31
20.	2.54	0.99	1.26	14.18
21.	4.90	0.97	0.53	14.54
22.	3.70	1.00	0.76	6.14
23.	4.05	0.99	0.79	14.50

Multpost for NO subgroup in all subjects

Table E17

Kolmogorov-Smirnov Tests for Matrix Albebra Multpost Items for All Subjects

NO Group				OK Group			
item	p	z	N	item	p	z	N
1)	Ø.9992	Ø.37ØØ	5	1)	Ø.ØØ67	1.4692	51
2)	1.ØØØØ	Ø/Ø	1	2)	Ø.7712	Ø.6633	55
3)	Ø.2Ø9Ø	1.Ø624	13	3)	Ø.7644	Ø.6675	43
4)	Ø.9459	Ø.5248	12	4)	Ø.9997	Ø.3494	44
5)	Ø.ØØØØ	5.3453	3	5)	Ø.12Ø1	1.1859	53
6)	1.ØØØØ	Ø/Ø	2	6)	Ø.272Ø	Ø.9982	54
7)	Ø.998Ø	Ø.3915	3Ø	7)	Ø.7921	Ø.65ØØ	26
8)	Ø.6979	Ø.7Ø8Ø	31	8)	Ø.8316	Ø.6236	25
9)	Ø.954Ø	Ø.5143	21	9)	Ø.9214	Ø.5514	35
1Ø)	Ø.9386	Ø.5333	24	1Ø)	Ø.8644	Ø.5999	32
11)	Ø.197Ø	1.Ø762	34	11)	Ø.2368	1.Ø325	22
12)	Ø.988Ø	Ø.4479	4Ø	12)	Ø.4371	Ø.8689	16
13)	Ø.4483	Ø.8613	35	13)	Ø.8973	Ø.5735	21
14)	Ø.86Ø2	Ø.6Ø3Ø	34	14)	Ø.8358	Ø.62Ø6	22
15)	Ø.6Ø73	Ø.7618	35	15)	Ø.6166	Ø.7563	21
16)	Ø.8174	Ø.6333	29	16)	Ø.98ØØ	Ø.4699	27
17)	Ø.2751	Ø.9952	49	17)	Ø.9Ø96	Ø.5626	7
18)	Ø.9833	Ø.4619	26	18)	Ø.997Ø	Ø.4Ø21	3Ø
19)	Ø.3216	Ø.9548	25	19)	Ø.55Ø1	Ø.7964	31
2Ø)	Ø.9676	Ø.4941	11	2Ø)	Ø.ØØØØ	3.5313	45
21)	Ø.1521	1.1349	31	21)	Ø.ØØ18	1.87Ø9	25
22)	Ø.ØØØØ	2.49Ø2	44	22)	Ø.Ø138	1.5772	12
23)	Ø.ØØØ7	1.9862	45	23)	Ø.9198	Ø.5529	11

Multpost for all subjects after 76 Fall semester ; goodness of fit test for Gamma distributions

Note. The following programs are used in connection with the programs listed in this table: "wb2," "wb2area," "kappa," "llab," "kolmo," "gamma," "wgraf," and "kgraf" programmed by Robert Baillie; "statedit," programmed by J. Michael Felty.

TABLE E18

Kolmogorov-Smirnov Tests for Matrix Algebra Test Items

OK Subgroup for Matinvtest

Weibull				Gamma		
item	p	z	N*	item	p	z
1)	.500	.827	27	1)	.002	1.88
2)	.723	.693	30	2)	.031	1.45
3)	.923	.550	28	3)	.980	.470
4)	.400	.897	27	4)	.376	.912
5)	.600	.768	27	5)	.008	1.658
6)	.440	.867	29	6)	.017	1.546
7)	.824	.629	23	7)	.915	.558
8)	.710	.702	27	8)	.214	1.057
9)	.737	.685	28	9)	.619	.755
10)	.292	.980	27	10)	.026	1.471
11)	.958	.510	20	11)	.722	.694
12)	.600	.766	24	12)	.006	1.710

* The total N is 30 and No subgroup was not analyzed.

Weibull parameters

item	t_o	c	μ_o
1)	18.66	.85	41.50
2)	8.78	.98	14.80
3)	2.88	2.29	24.35
4)	7.51	1.42	27.35
5)	7.94	0.74	14.68
6)	7.87	0.91	12.50
7)	1.61	1.84	20.79
8)	13.50	.93	36.98
9)	6.25	1.41	32.21
10)	5.74	.93	21.23
11)	15.37	1.03	21.69
12)	4.92	.63	41.25

TABLE E19

Kolmogorov-Smirnov Tests for Matrix Algebra Test Items

OK Subgroup for Transtest

Weibull				Gamma		
item	p	z	N*	item	p	z
1)	.464	.851	16	1)	.579	.778
2)	.694	.710	26	2)	.073	1.285
3)	.146	1.144	34	3)	.011	1.618
4)	.941	.530	11	4)	.874	.592
5)	.950	.519	26	5)	.896	.574
6)	.061	1.321	29	6)	.062	1.316
7)	.216	1.054	34	7)	.000	2.304

* The total N is 38

OK Subgroup for Eigtest

Weibull				Gamma		
item	p	z	N*	item	p	z
1)	.971	.488	20	1)	.513	.819
2)	.626	.751	39	2)	.801	.644
3)	.829	.625	44	3)	.825	.628
4)	.918	.555	28	4)	.388	.904
5)	.666	.727	34	5)	.367	.919
6)	.999	.285	16	6)	.978	.475
7)	.609	.761	25	7)	.701	.706
8)	.727	.691	30	8)	.640	.742

* The total N is 56

TABLE E20

Kolmogorov-Smirnov Tests for Matrix Algebra Test Items

NO subgroup for Transtest

Weibull				Gamma		
item	p	z	N*	item	p	z
1)	.965	.498	22	1)	.718	.696
2)	.872	.594	12	2)	.544	.800
3)	.997	.399	4	3)	.945	.526
4)	.038	1.407	27	4)	.123	1.182
5)	.829	.625	12	5)	.444	.865
6)	.732	.688	9	6)	.116	1.193
7)	.964	.500	4	7)	.885	.584

* The total N is 38

No subgroup for Eigtest

Weibull				Gamma		
item	p	z	N*	item	p	z
1)	.4665	.8492	36	1)	.477	.843
2)	.9801	.4697	17	2)	.838	.619
3)	.9993	.3649	12	3)	.946	.525
4)	.5266	.8108	28	4)	.587	.774
5)	.7481	.6776	22	5)	.832	.623
6)	.9518	.5173	40	6)	.984	.461
7)	.7674	.6656	29	7)	.491	.834
8)	.9323	.5402	24	8)	.959	.507

* The total N is 56

TABLE E21

Kolmogorov-Smirnov Tests for Matrix Algebra Test Items

Weibull Parameters for Trantest and Eigtest Items

OK subgroup for Transtest			
item	t_o	c	μ_o
1)	10.63	1.12	31.65
2)	15.45	.84	67.16
3)	3.80	1.01	13.99
4)	6.15	1.24	22.53
5)	5.99	1.23	25.11
6)	2.86	.94	10.25
7)	2.93	.91	7.36

No subgroup for Transtest			
item	t_o	c	μ_o
1)	10.35	.74	95.10
2)	4.09	.80	104.27
3)	9.33	.84	7.40
4)	2.03	1.17	33.58
5)	2.84	1.11	20.75
6)	5.94	.57	9.73
7)	4.36	.94	5.36

OK subgroup for Eigtest			
item	t_o	c	μ_o
1)	10.68	1.08	75.94
2)	3.28	1.63	41.60
3)	6.88	1.43	27.26
4)	5.43	1.19	17.42
5)	7.69	1.04	17.54
6)	14.09	1.09	22.01
7)	6.37	1.27	58.50
8)	8.26	1.04	35.83

NO subgroup for Eigtest			
item	t_o	c	μ_o
1)	3.29	1.75	91.00
2)	15.09	1.07	20.50
3)	11.22	1.06	14.50
4)	4.03	1.44	16.98
5)	3.86	1.16	25.81
6)	1.66	1.39	53.65
7)	8.03	1.13	39.00
8)	1.57	1.23	34.51

Appendix F

Graphs of Conditional Response Rate

	c	t_0	μ_0
item 3 :	1.049	6.5019	19.29
item 4 :	1.022	7.6480	14.47

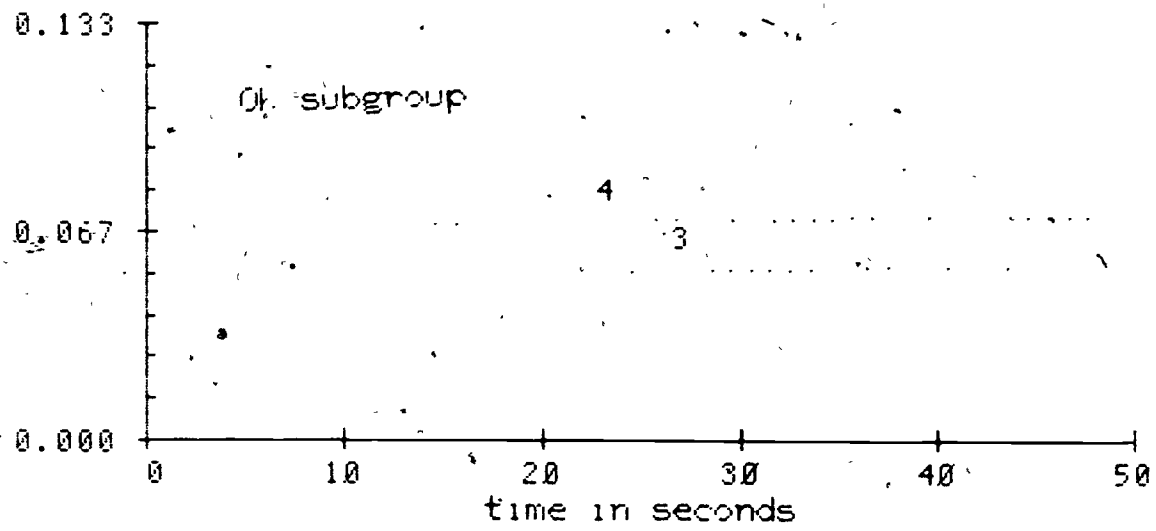


Figure F1 Comparison of conditional response rates of items 3 and 4 for OK subgroup

	c	t_0	μ_0
item 5 :	2.103	5.6420	17.05
item 6 :	0.8525	0.0001	10.79

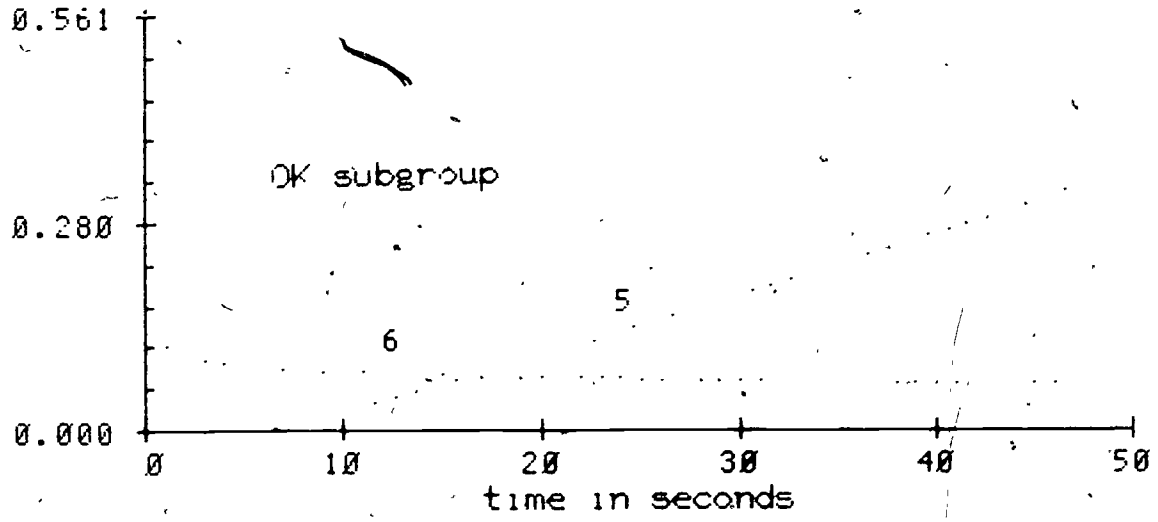


Figure F2 Comparison of conditional response rates of items 5 and 6 for OK subgroup

		c	t_0	μ_0
item	3 :	1.240	6.0210	23.34
item	4 :	2.188	0.0001	32.35

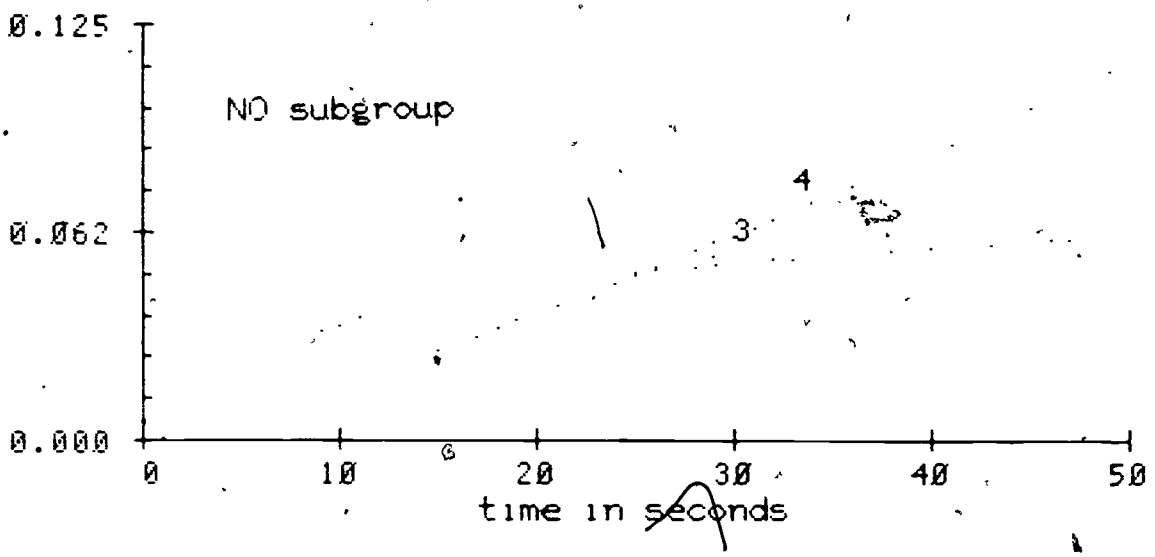


Figure F3 Comparison of conditional response rates of items 3 and 4 for NO subgroup

		c	t_0	μ_0
item	5 :	1.251	7.5420	13.70
item	6 :	1.353	5.5105	12.67

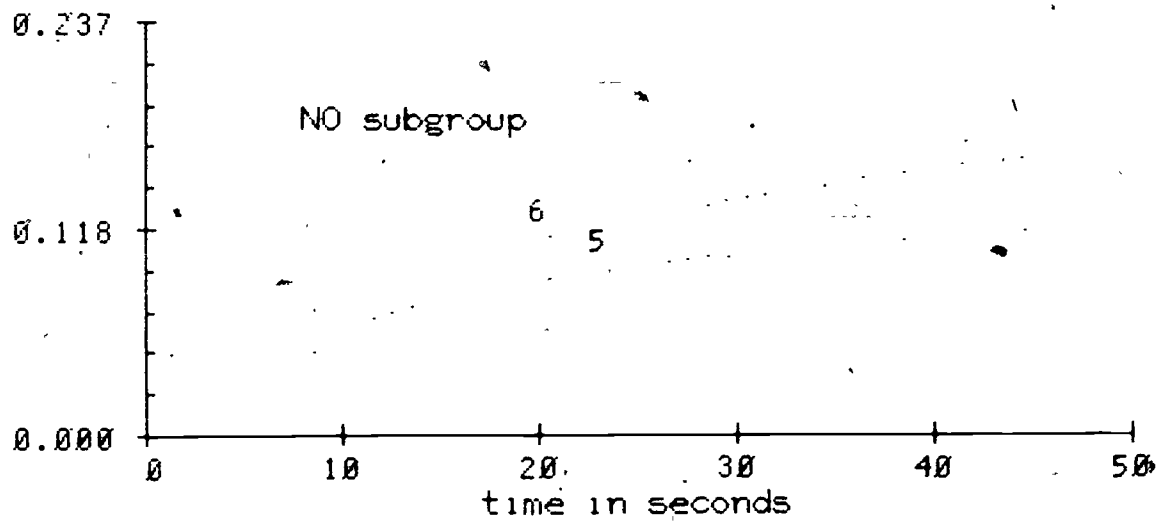


Figure F4 Comparison of conditional response rates of items 5 and 6 for NO subgroup