DOCUMENT RESUME

ED 154 137 CE 015 211

AUTHOR Bunch, Michael B.
TITLE Making Decisions about Adult Learners Based on Performances on Functional Competency Measures.
PUB DATE 6 Apr 78
NOTE 40p.; Paper presented at the Annual Adult Education Research Conference (San Antonio, Texas, April 5-7, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Adults; *Basic Skills; *Criterion Referenced Tests; *Cutting Scores; Measurement Techniques; Performance Criteria; Performance Tests; *Test Reliability; *Test Validity
IDENTIFIERS Adult Performance Level

ABSTRACT

The validity and dependability of functional competency tests for adults are examined as they relate to the information needs of instructional decision makers. Test data from the Adult Performance Level (APL) Program (funded by the U.S. Office of Education at the University of Texas at Austin) is used to illustrate key points. In the discussion of validity, the importance of a test's demonstrated relevance to functional competency is discussed in terms of the definitions of the competency. Issues of content vs. criterion validity are examined particularly with reference to the APL study. Some of the problems inherent in setting and applying cutoffs (points on a scale of scores which define levels of competence) are then discussed, and the author reviews several procedures to aid in setting and adjusting cutoffs (those used by Nedelsky and by Emrick, and Bayesian techniques used by Northcutt). In the discussion of dependability (the degree to which scores are replicable) the author reviews briefly the work of Bob Brennan and Mike Kane (based on that of Cronbach and others) in the area of defining and assessing psychometric properties of criterion-referenced tests. In conclusion it is pointed out that the instructional decision maker may raise or lower a cutoff as information justifies such action but that there will be instances in which trade-offs between dependability and validity may become necessary. (JT)

MAKING DECISIONS ABOUT ADULT LEARNERS BASED ON PERFORMANCES ON FUNCTIONAL COMPETENCY MEASURES

Michael B. Bunch
NTS Research Corporation
Durham, North Carolina

Presented at the Annual Meeting of the
Adult Education Research Conference,
San Antonio, Texas, April 6, 1978

MAKING DECISIONS ABOUT ADULT LEARNERS BASED ON
PERFORMANCES ON FUNCTIONAL COMPETENCY MEASURES

Adult Basic Education (ABE) has long concerned itself with those individuals whose ability to function within society is at a marginal level. A symptom of the condition of marginal functioning has always been either illiteracy or functional illiteracy. Currently the phrase "functional competency" is perhaps more comprehensive. Adult educators have risen to the challenge of educating adults to be functionally competent, and the concept of functional competency has gained national recognition.

The Economic Opportunity Act of 1964 (PL 88-452, Title IIB) and the Adult Education Act of 1966 (PL 89-750, Title III) have focused national attention on the functional competency needs of adults. A national "Right to Read" Adult Movement (sponsored by the U.S. Department of Health, Education, and Welfare) adopted the following policy statement in 1970:

> The challenge is to foster through every means the ability to read, write, and compute with the functional competence needed for meeting the requirements of adult living[1].

This focus on functional competencies, on "coping skills", eventually led to the U.S. Office of Education funded Adult Performance Level study at the University of Texas at Austin. The purpose of the study was twofold; to specify the competencies required for functioning in society, and to develop devices for assessing those competencies. The underlying assumptions were, of course, that definable competencies did exist and that they could be measured.

Functional competency, when operationally defined in terms of specific tests, typically implies that there is a cutoff point or set of cutoff points which define levels of competence. In the case of one cutoff point, those persons scoring at or above the cutoff are considered competent, while those scoring below are not. In the case of two or more cutoff points, individuals are placed into categories as a result of their scores in relation to the various cutoffs.

The decision maker is immediately faced with two questions; these concern the validity of the test and the degree to which scores are replicable. For the purpose of this paper, these concerns will be referred to as validity and dependability. The remainder of this paper will be devoted to the issues of validity and dependability of measurement as they relate to information needs of instructional decision makers. Test data from the Adult Performance Level Program (ACT, 1976, 1977) will be used to illustrate key points.

Validity

Decision makers may place several requirements on tests of functional competency. These tests must, above all, have some demonstrated relevance to functional competency, as defined in a way acceptable to the decision maker. Thus, for example, if functional competency is defined in terms of social and economic success (and if this definition is acceptable to the decision maker) then tests of functional competency must demonstrate a positive correlation with measures of social and economic success in order to be considered valid (i.e., to possess criterion validity). If, on the other hand, competency is defined strictly in terms of mastery of a specified set of objectives then the validity of functional competency tests rests in the judged relevance of individual items to the several objectives (content validity). In any event, the operational definition used in the construction of a competency measure (and the definition may very well suggest both content and criterion validity) will dictate validation procedures to a certain extent. Whether the decision maker uses a locally constructed measure, or a nationally standardized one, the relationship between the acceptable definition of competency and the available validity data should be examined carefully.

Nafziger, Thompson, Hiscox, and Owen (1975) reviewed several measures of what they termed "functional literacy" (for all practical purposes very similar to functional competency but less comprehensive). Of the four criterion referenced tests reviewed, all were rated as good with respect to content or construct validity and fair to poor with respect to criterion validity. Overall, the validity of each measure (including the 42 item Texas APL Survey) was rated as fair. It is clear, however, that the developers of the four tests concentrated on content validity, while the definition accepted by Nafziger et al. included both content and criterion validity.

The definition of functional competency developed by the University of Texas APL research team (Northcutt, Selz, Shelton, & Nyer, 1975) stated that: 1) the term functional competency is meaningful only in a specific societal context; 2) functional competency is best described as the application of a set of skills to a set of general knowledge areas; 3) functional competency results from a combination of individual capabilities and societal requirements; and 4) functional competency is directly related to success in adult life. Points (2) and (4) of the definition may be viewed as dictating content and criterion validation procedures. Yet, Northcutt _et al_. seemed to concentrate on point (2), in terms of validity information, in their final report. This emphasis is reflected in the fact that Nafziger _et al_. rated the APL Survey very highly in terms of content validity and very poorly in terms of criterion validity.

A criticism on similar grounds was later voiced by Griffith and Cervero (1977). They argued that both the original University of Texas APL researchers and American College Testing Program APL staff had devoted too little attention to criterion validity. More recently, Cervero has provided some criterion validity information regarding the APL Survey[2]. In a reanalysis of original APL Survey data, Cervero found significant correlations between Texas developed APL Survey scores and measures of success. These were .56 for years of schooling, .33 for occupational status, and .39 for family income. All correlations were based on 5,000 to 8,000 responses and significant beyond the .001 level. According to Cervero (p.4), "Since the correlations between APL test score and indicators of 'success' are about as good as would be expected, it could be argued that the APL test is directly related to 'success' in adult life, as the developers assume".

Correlations between APL Content Area Measures and adult success criterion variables were not as high as those found for the original APL Survey. These correlations, reported in the APL Content Area Measure Technical Supplement, (ACT, 1977f) ranged from .09 to .19 for family income (median r = .15) and from .19 to .21 for years of education (median r = .20). All correlations were based on 650 to 1,100 responses. Although all were significant, they were less than one might expect, given previous findings (e.g. Jencks _et al._, 1972).

Performance on APL Content Area Measures is understandably interpreted in terms of instructional goals. Whereas levels on the original APL Survey (Northcutt _et al._, 1975) were couched in terms of likelihood of success in adult life, ACT level definitions are as follows:

Level 1 - Has an inadequate degree of competency - a definite need for study and remediation to meet the APL goals and objectives through the application of basic skills.

Level 2 - Has a marginal degree of competency a need for study and review to meet the APL goals and objectives through the application of basic skills.

Level 3 - Has an adequate degree of competency - may need some review to continue to meet the APL goals and objectives through the application of basic skills.

Given these definitions, the instructional decision maker has no basis for relating test performance directly to likelihood of success in life. Learners are evaluated strictly in terms of objective mastery.

A question which immediately arises when adjectives such as "inadequate", "marginal", or "adequate" are used, no matter what the context, is "By what criterion?" That is, what is the standard by which these labels are attached to individual performances? There is a score, for example, below which performance is judged to be inadequate and above which performance is judged to be adequate (or marginal). The process by which these scores are established is of crucial importance. Analysis of this process is no less important than an analysis of the content or criterion validity of the test because effects of the process on the learner are no less profound than those of test validity.

Greater attention will be paid to the setting of cutoffs within the section on dependability but it seems important to outline here some of the problems inherent in setting cutoffs and some of the related problems faced by instructional decision makers. It is perhaps little consolation to find that these problems are not unique to the field of functional literacy/competency. They are simply a little more actue because of the current visibility of functional competency.

It is typically the case that criteria or cutoff scores are set more or less arbitrarily[3]. This is true even of many nationally published tests which have cutoffs. An excellent review of some of the procedures by which cutoffs may be set more objectively may be found in an article by John Meskauskas (1976). Although there is a certain degree of arbitrariness in all procedures reviewed, elements of objectivity are introduced which have the effect of reducing arbitrariness, to varying degrees, in each of the methods. Two procedures may serve as illustration, although others are certainly possible and defensible.

The Minimum Pass Level (MPL) developed by Nedelsky (1954) utilizes the judgements of several persons who rate individual items with respect to difficulty. Let us assume that seven instructors (A through G) each rate one hundred test items (1 through 100). Instructor A looks at item 1 and predicts the chances of the hypotethically lowest passing learner (i.e., the least competent of the competent) for answering the item correctly. Instructor A then does the same with items 2 through 100 and adds the probabilities to get an MPL. Instructors B through G do the same. One can then express the minimum passing level (MPL) as follows:

$$MPL = \overline{M}_{FD} + K\sigma_{FD} \qquad (1)$$

where $\overline{M}_{FD}$ is the mean of the individual instructor MPLs and σ_{FD} is the standard deviation of the distribution of individual MPLs. FD refers to a cutoff between grades of F and D. For tests such as the APL Survey or Content Area Measures, one might just as easily focus on the cutoff separating levels 1 and 2 and on the cutoff separating levels 2 and 3. K is a constant which may be adjusted to control the percentage of marginal students who "pass" the test. The essential subjective elements are the individual predictions of learner success on given items and the setting of the value of K. This method does have some advantages over a totally <u>ad hoc</u> approach in that it does focus on individual items and forces some structure onto the process. Ebel (1972) has developed a similar procedure which essentially extends Nedelsky's model into two dimensions (relevance and difficulty).

A procedure attributed to Emrick (1971) draws upon decision theory in that the test designer or administrator must express certain subjective factors upon which he or she bases decisions. Although the procedure treats competency as an all or none trait (i.e., there is no underlying continuum of mastery; a learner has either mastered or failed to master a given curriculum). It may be viewed as helpful in setting cutoffs because it relates test performance to performance in other areas and is best applied at the subtest level (i.e., units of about ten items). The decision maker is forced to make a statement about how bad different kinds of errors of classification would be. Let us call the erroneous placement of a non-master into the master category (on the basis of a response to any given item) a Type 1 error (false positive) and the converse error a Type 2 error (false negative). The probability of making a Type 1 error will be expressed as α, while the probability of a Type 2 error will be expressed as β. Now the decision maker must express in a ratio the relative losses associated with these two types of errors. Emrick (1971) calls this the ratio of regret (RR). This ratio is purely subjective unless, of course, real costs may be determined for each type of loss. The optimal cutting score (C) may be expressed in terms of test length (n) and these other factors as follows:

$$C = \frac{\log \frac{\beta}{1 - \alpha} + 1/n\ (\log RR)}{\log \frac{\alpha\beta}{(1 - \alpha)(1 - \beta)}} \qquad (2)$$

Information about learners accumulated over a period of time may provide empirical estimates of α and β in equation (2). If, for example, it is discovered that five percent of those learners who answer certain items correctly have not actually mastered the content, then $\alpha = .05$. If, on the other hand, ten percent of learners who respond incorrectly to certain items are actually masters, then $\beta = .10$. Assuming now that the two types of errors are equally serious, RR would equal 1.0. Thus, for a 10 item sub-test equation (2) would yield a cutting score of 4.4 which could be rounded off to 4 or 5. The values of C for a whole test could be added together to yield a total test cutoff. In the special case where Type 1 and Type 2 errors have an equal probability of occurring (α=β), and both types are considered equally serious (RR = 1.0) it can be shown that the cutoff score will always be exactly half the total number of items.

Of course, it will not always be the case that all things will be equal, and the cutoff will have to be set at some point other than 0.5. Figures 1 through 3 are provided to show what happens to C as each of the parameters changes. As can be seen from Figure 1, the value of C levels off very quickly as RR increases for the given values of α and β. In other words, the value of

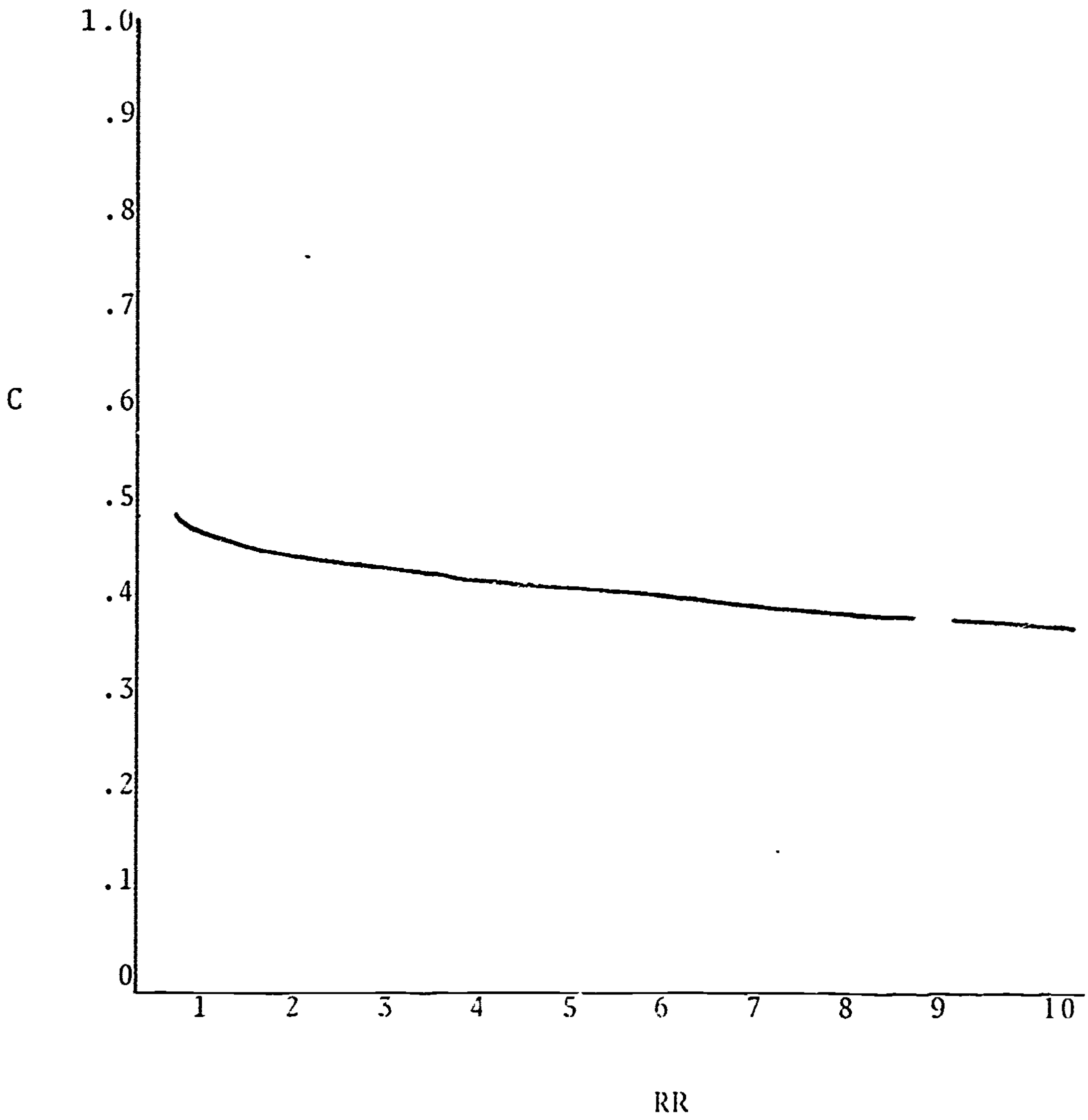

Figure I. Cutoff (C) as a function of ratio of regret (RR) with values of α and β fixed at .05 and .10, respectively.

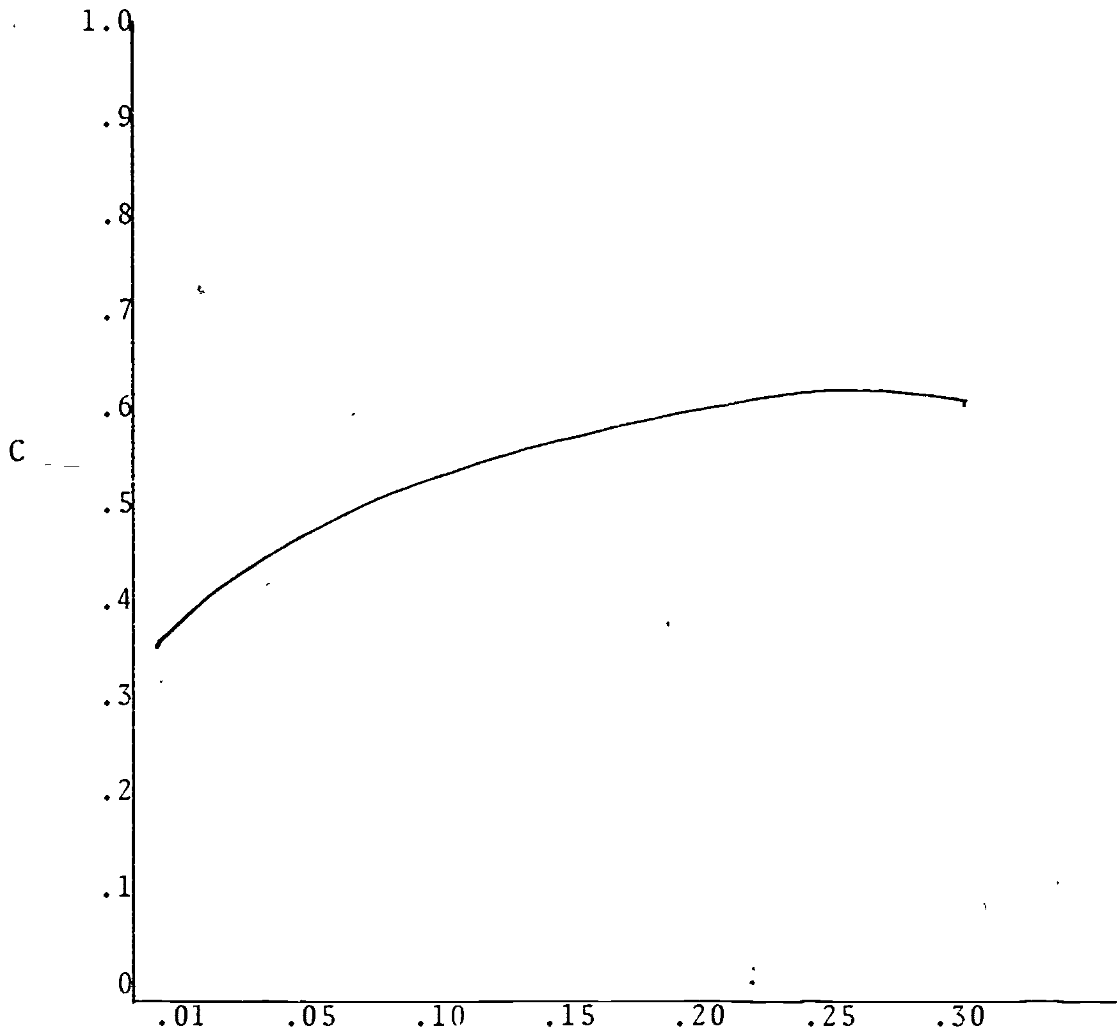

Figure 2. Cutoff (C) as a function of probability of false positive error (α) with values of β and RR fixed at .10 and 1.0 respectively.

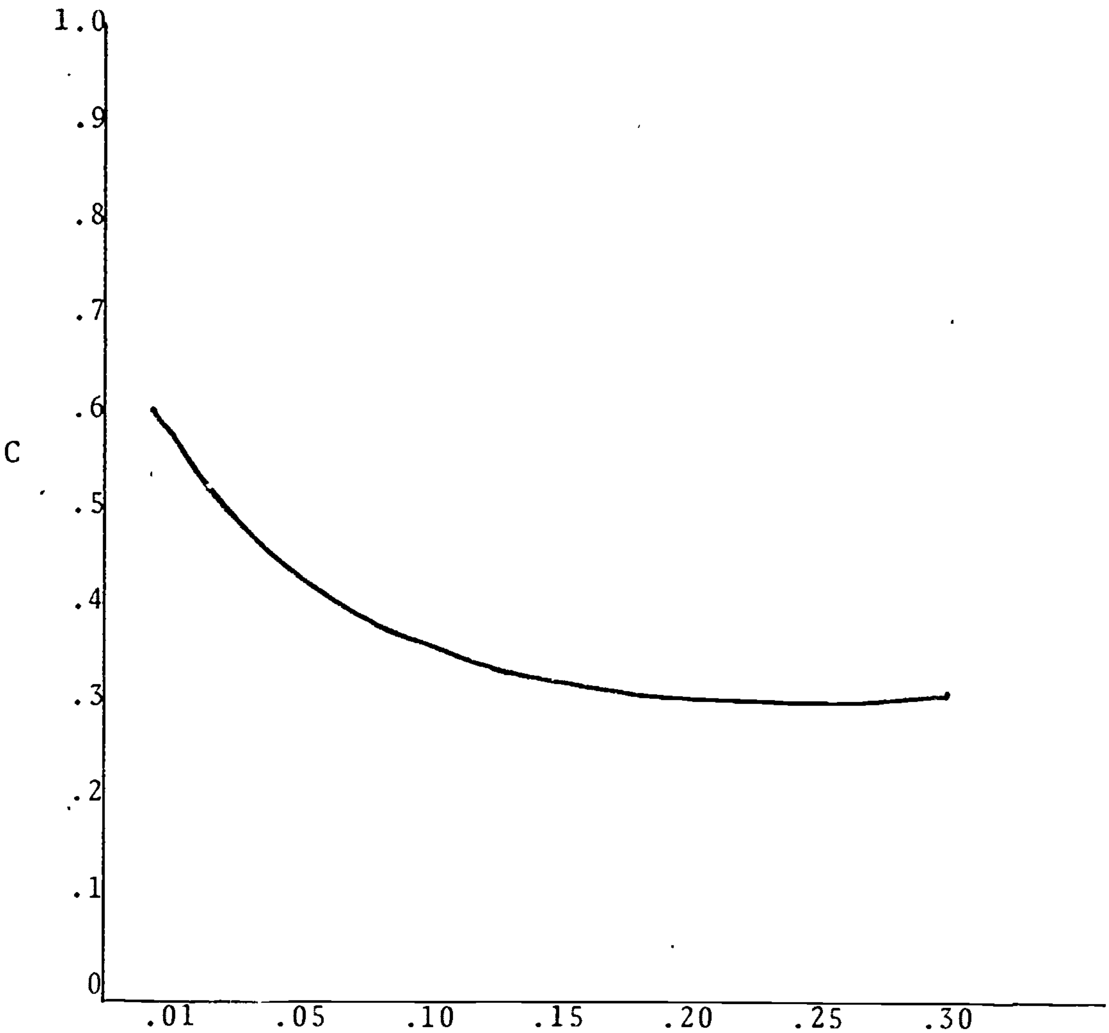

Figure 3. Cutoff (C) as a function of probability of false negative error (β) with values of α and β fixed at .05 and .10, respectively.

the most subjective parameter of equation (2) seems to have little impact on C for these data. Although the largest value of RR is 100 times as great as the smallest value, the range is only .09 (i.e., from .39 to .48).

On the other hand, values of C change rather dramatically as either α and β increases. In Figure 2 the value of C ranges from .34 to .62 while α goes from .01 to .30. The range of C is thus three times that of C in Figure 1. Likewise, in Figure 3, the range of C is from .32 to .60 or about three times the range of C in Figure 1. Also note that as α increases, C increases, while C decreases as both β and RR increase. As the likelihood of classifying non-masters as masters increases, one is forced to raise the cutoff. As the likelihood of classifying masters as non-masters increases, one is forced to lower the cutoff. Similarly, if the second type of misclassification is considered to be a more serious mistake (larger regret) than a misclassification of the first type (smaller regret), then it will be necessary to lower the cutoff. Although other values for each of the three parameters could have been chosen, these are representative of likely values one might obtain empirically. Other sets of parameters might yield very different kinds of curves. In fact for some values of α and β, C will be undefined, for example, when $\alpha + \beta = 1.0$, or when all examinees are misclassified. Under such conditions, the decision maker is well advised to choose an alternative method for establishing cutoffs.

The point of this admittedly rather lengthy discourse is this: the setting of cutoffs on functional competency measures need not be completely arbitrary. In fact, because behavioral manifestations of competency will vary from place to place, it is advisable to consider setting one's own population specific cutoff. The instructional decision maker can and should maintain a constant surveillance over the effects of cutoffs on placement and subsequent performance of learners and adjust as he or she sees need to do so. This adjustment becomes easier if the criterion is something with which the decision maker is quite familiar, such as curriculum objectives. This adjustment becomes more difficult if the criterion is something with which the decision maker is less familiar, such as the actual life success of individual learners. This reason, as well as for other reasons, it would seem more appropriate for adult educators to concentrate on curriculum objectives rather than on global indicators of life success. While several procedures are available to aid in setting cutoffs, the decision maker should rely on the method which matches his or her definition of competency and characteristics of the program and learners.

A procedure unlike either of the two just described (*viz.*, Nedelsky, 1954; Emrick, 1971) was used by Northcutt (1974) to set cutoffs on the APL Survey. In his procedure, Northcutt used Bayesian techniques (see, for example, Novick, 1973, for a review of Bayesian applications). First, he obtained a rough concensus regarding the operational definition of adult success. Next, the Opinion Research Corporation was employed to conduct a nationwide survey of a representative sample of adults to estimate the percentages of adults classified at each success level. This same sample was also given the first version of the APL Survey. It was found that items could discriminate among the three groups of adults (with respect to life success). The test score related level classifications which ultimately emerged took into account this discriminating power of items. The process underwent several refinements before the final cutoffs were set. By this process, it was estimated that roughly 20% of the adult population of the United States were functionally incompetent (Level 1), 34% were marginally competent (Level 2), and 46% were proficient (Level 3).

More recently, Jerry Williams[4] set cutoffs on an APL test by comparing the performances of various groups of adults on the test. These various subgroups were aggregated into two major groups, productive and marginally productive. The productive group contained professionals, machinists, craftsmen, sales workers, farmers, and so on. The marginally productive group consisted of prison inmates, unemployed, and persons for whom English was not a native tongue (but who were receiving English instruction). By comparing the median scores for all groups, Williams found a fairly clean break at about 70%. This percentage was taken as a rough estimate of a desired level of performance. The actual cutoff used was moderated by a procedure similar to Emrick's (1971) such that the actual cutoff was .60.

The examples just given show the relationship between test validity and setting of cutoffs. In one case (Nedelsky, 1954) the setting of a cutoff was related more or less to content validity. In the other cases, cutoffs were more clearly related to criterion validity. The key issue here is that the subjectivity of classification of learners may be greatly reduced through a modicum of effort. Given the context of validity based cutoffs (which need not be elaborately worked out), the instructor of adult learners may render very defensible, data based judgements.

Dependability of Measurement

A specific implication of functional competency testing is that adults are not ranked in order of score but rather that each person's score is compared to a predetermined cut-off or set of cut-offs. Thus, functional competency testing is typically outside the realm of norm-referenced testing and well within the realm of criterion, or domain referenced testing.

Most of test theory, as we know it today, has been developed around the concept of ranking individuals along some continuum. The concept of cut-off, or minimum level of performance has never been very important. Within the past two decades, however, this concept has become very important. The individualized instruction movement of the late 1940's and beyond raised many technical questions, including a number related to testing. These questions were addressed by several researchers from about 1960 to the present. Most of the research focused on individual items; how to construct them, how to select them, etc. A few researchers concentrated on assessing the characteristics of decision making procedures, which included total test qualities as well as the setting of cutoffs.

The most promising work in the area of defining and assessing psychometric properties of criterion-referenced tests has been done by Bob Brennan and Mike Kane (Brennan, 1977a, 1977b; Brennan

& Kane, 1977, in press; Kane & Brennan, 1977). Their work stems directly from that of Cronbach, Gleser, Nanda, & Rajaratnam (1972). Whereas the work of Cronbach et al. concentrated on norm-referenced tests, Brennan and Kane have focused on criterion or domain-referenced tests. One difference in the two approaches lies in the fact that Brennan and Kane allow for cut-off scores.

While I will attempt to summarize these works here enough to shed some light on the remainder of the paper, this review is by no means exhaustive or comprehensive. Those interested are directed especially to the book by Cronbach et al. (1972) and the article by Brennan & Kane (1977). Following this review, I shall present data from the development of the APL Content Area Measures (ACT, 1977a-f) which illustrate uses of dependability/generalizability theory. I shall also attempt to demonstrate the applicability of such procedures to local decision-making processes involving adult learners and measures of functional competency.

Cronbach et al. (1972) suggested a liberalization of test theory to take into account more than two facets in the determination of the reliability of measures. This liberalization has come to be known as generalizability theory, as opposed to classical test theory. While classical test theory treats reliability as the ratio of two variances (cf. Guilford, 1954; Lord & Novick, 1968), this approach considers only two types of variance; namely true score and error. In classical terms, observed score variance, $\sigma^2(t)$ is viewed as divisible into two components as defined in the following equation:

$$\sigma^2(t) = \sigma^2(T) + \sigma^2(e), \qquad (3)$$

where $\sigma^2(T)$ is true score variance, and $\sigma^2(e)$ is error variance. In this context reliability (r) is expressed as a ratio:

$$r = \frac{\sigma^2(T)}{\sigma^2(T) + \sigma^2(e)} \tag{4}$$

In the most straightforward case, a group of examinees is given a set of items, and this process is called a test administration. In this simplest case, at least three definable things or components enter into total score variance. These are the items, the examinees. and error. In the terminology of Cronbach et al. the observed score of examinee p on item i (X_{pi}) may be expressed as

$$X_{pi} = \mu + \pi_p + \beta_i + \pi\beta_{pi} + e \tag{5}$$

where μ is the grand mean across persons and items; π_p is the effect due to person p; β_i is the effect due to item i; $\pi\beta_{pi}$ is the effect due to the interaction of person p and item i; and e is experimental error. Since person p only takes item i once, it is not possible in this situation to estimate the interaction effect. Therefore, the effects $\pi\beta_{pi}$ and e are lumped together in a common error term. Thus,

$$X_{pi} = \mu + \pi_p + \beta_i + \pi\beta,e \tag{6}$$

where $\pi\beta,e$ is the common error term, and all other terms are as defined in equation (5).

Reliability, within the context of generalizability theory, is also expressed in terms of variances or variance components. However, before entering into a discussion of these components of variance, it will be necessary to discuss two contexts within which variance components are computed. These contexts are generalizability studies and decision studies.

Cronbach et al. (1972) distinguish between generalizability studies, or G-studies, and decision studies, or D-studies. In a G-study, one is typically interested primarily in a theoretically infinite population of examinees and universe of items. In a D study, one is typically interested in a more narrowly defined group of examinees and/or items. A test publisher may, for example, administer a new test to a nationally selected group of examinees. The intent of this administration may be to accumulate information about the degree to which the test results generalize to the domain (or item universe) of interest. In a D-study a local decision maker may be interested only in the performance of a specific group of examinees (a class) on a specific set of items (a form of the test). Note, however, that the test developer may also wish to conduct a D-study using all or part of the information gathered in the G-study.

Once a test has been administered, it is possible to view the results in terms of a two facet analysis of variance problem where the facets are persons (p) and items (i). In this p-by-i design, the score of person p on item i may be expressed as in equation (6). By using analysis of variance procedures, it is possible to obtain mean squares (MS) due to persons, items, and the person-item interaction, which will be taken as the error component. It can be shown (cf. Brennan, 1977a) that variance components are directly estimable from mean squares. Specifically,

$$\hat{\sigma}^2\ (p) = \{MS\ (p) - MS\ (pi)\}/n_i, \tag{7}$$

$$\hat{\sigma}^2\ (i) = \{MS\ (p) - MS\ (pi)\}/n_p, \tag{8}$$

and

$$\hat{\sigma}^2\ (pi) = MS\ (pi), \tag{9}$$

where MS (p) is equal to the mean square for persons, MS (i) is equal to the mean square for items, MS (pi) is equal to the mean square for the person-by-item interaction; $\hat{\sigma}^2$ (p), $\hat{\sigma}^2$ (i), and $\hat{\sigma}^2$ (pi) are the estimated G-study variance components for persons, items, and the interaction term, respectively.

These estimates represent the variance components obtained in the simplest case; i.e., the person-by-item case. Far more complex cases are possible (and are treated by Brennan, 1977a) but need not be examined here. These variance component estimates are quite helpful to the test consumer in terms of evaluating various test of similar content. In fact, the American Psychological Association (APA) American Educational Research Association (AERA) and National Council on Measurement in Education (NCME) strongly suggest reporting G-study variance components along with reliability data in technical manuals for published tests (APA, 1974).

D-study variance components may be derived directly from G-study components, once the testing model has been defined and a decision has been made as to how far one wants to generalize results. Brennan (1977a) has devised a system to aid the decision maker in specifying these parameters and deriving variance components.

For the purpose of this paper, let us assume that we are interested in being able to generalize over a potentially infinite universe of items. In this case, the D-study variance components may be expressed as follows:

$$\hat{\sigma}^2(p) = \hat{\sigma}^2\ (p), \tag{10}$$

$$\hat{\sigma}^2\ (I) = \hat{\sigma}^2\ (i)/n'_i, \tag{11}$$

and

$$\hat{\sigma}^2\ (pI) = \hat{\sigma}^2\ (pi)/n'_i. \tag{12}$$

In equation (10), the D-study variance component for persons is equal to the G-study variance component for persons. This will be the case when person is the unit of analysis (other possibilities for unit of analysis include class, school, state, etc.). In equations (11) and (12), the capital I denotes sampling across items. The term n'_i in equations (11) and (12) represents the number of items in the particular test used in the D-study.

Given these D-study variance components, it is possible to estimate two types of error for a given test. One is associated with norm referenced testing situations and is denoted $\hat{\sigma}^2(\delta)$. The other is associated primarily with criterion referenced testing and is denoted $\hat{\sigma}^2(\Delta)$. Cronbach et al. (1972) indicate that $\hat{\sigma}^2(\delta)$ is appropriate for expressing error in terms of the deviation from the population mean. $\hat{\sigma}^2(\Delta)$ is, on the other hand, appropriate for expressing error associated with the differences between a given examinees' item universe scores and observed scores. In terms of equations (11) and (12), we may operationally define $\hat{\sigma}^2(\delta)$ and $\hat{\sigma}^2(\Delta)$ as follows:

$$\hat{\sigma}^2(\delta) = \hat{\sigma}^2(pI), \tag{13}$$

and

$$\hat{\sigma}^2(\Delta) = \hat{\sigma}^2(I) + \sigma^2(pI) \tag{14}$$

where all terms are as defined above and in equations (11) and (12).

Cronbach et al. (1972) use the term $\hat{\sigma}^2(\delta)$ in the calculation of the generalizability index, $\varepsilon\hat{\rho}^2$ or the ratio of universe score variance to expected observed score variance. This is essentially the same coefficient as coefficient alpha (Cronbach, 1951) and KR-20 (Kuder & Richardson, 1937). It is traditionally taken as the estimate of the internal consistency reliability of a test and may be expressed as

$$\varepsilon\rho^2 = \frac{\hat{\sigma}^2(p)}{\hat{\sigma}^2(p) + \hat{\sigma}^2(pI)} \tag{15}$$

where all terms are as defined above and in equations (10) and (12).

Brennan & Kane (1977) used the error term $\hat{\sigma}^2(\Delta)$ in developing an index of dependability for criterion referenced tests or any test which contains one or more cut-offs. Their index, called $\underline{M}$ ($\underline{C}$) may be expressed in terms of variance components as follows:

$$\underline{M}\ (\underline{C}) = \frac{\hat{\sigma}^2 p + (\mu - C)^2}{\hat{\sigma}^2 p + (\mu - C)^2 + \hat{\sigma}^2(I) + \hat{\sigma}^2(pI)} \tag{16}$$

where μ is the population score mean, C is the cut-off score, and other terms are as defined in equations (10) through (12). When items are scored simply as correct/incorrect (or 1/0), Brennan & Kane (1977) have shown that equation (16) may be estimated from sample means and variances:

$$\widehat{\underline{M}\ (\underline{C})} = 1 - \left[\frac{1}{n_i - 1}\right] \left[\frac{X_{PI}\ (1 - X_{PI}) - S^2(X_{pI})}{(X_{PI} - C)^2 + S^2\ (X_{pI})}\right] \qquad (17)$$

where X_{PI} is the sample mean over items and examinees, $S^2(X_{pI})$ is the sample variance of persons' scores over items, and $\underline{M}\ (\underline{C})$ stands for the estimated value of $\underline{M}\ (\underline{C})$.

Finally, when the cut-off is equal to the sample mean ($C=X_{PI}$), Brennan (1977b) has shown that:

$$\widehat{\underline{M}\ (\underline{C})} = 1 - \left[\frac{1}{n_i - 1}\right] \left[\frac{X_{PI}\ (1 - X_{PI}) - S^2\ (X_{pI})}{S^2\ (X_{pI})}\right] \qquad (18)$$

where all terms are as defined in equation (17). This equation is identical to the internal consistency estimate of tests derived by Kuder & Richardson (1937) in their formula 21. This value is the lowest possible value of $\widehat{\underline{M}\ (\underline{C})}$ for a given testing situation and will be denoted KR-21 throughout the remainder of this paper. It can also be shown that as the value of C approaches the maximum or minimum possible score, $\widehat{\underline{M}\ (\underline{C})}$ will approach its maximum value, and as C approaches X_{PI}, $\widehat{\underline{M}\ (\underline{C})}$ approaches KR-21. Implications for the setting of cut-offs are discussed in the following example.

Data from the development of the APL Content Area Measures (ACT, 1977 a-f) are used here because of the relevance of the APL program to functional competency and because generalizability/dependability procedures were used in their development. Data were collected in the spring (April) of 1977 from a total of 4,563 adult education students representing a cross section of four regions and five different community sizes in the United States. Inasmuch as there were five Content Area Measures, each adult education student responded to items in only one content area. Table 1 shows the number of items in each Content Area Measure (CAM) and the number of examinees associated with the development of each CAM.

Table 1

Numbers of Items and Examinees Associated with each Content Area Measure

Content Area Measure	Items	Examinees
Community Resources	51	855
Occupational Knowledge	42	866
Consumer Economics	66	1,148
Health	45	841
Government and Law	45	853

Variance components for each test were estimated through multiple matrix sampling procedures (Shoemaker, 1973). These variance components were then used to obtain values of $\hat{\sigma}^2(\delta)$, $\hat{\sigma}^2(\Delta)$, $\varepsilon\hat{\rho}^2$, and $\widehat{M(C)}$. Since each CAM has, in effect, two cut-offs, two values of $\widehat{M(C)}$ were calculated for each test. In addition, other values of $\widehat{M(C)}$ were obtained for a range of cut-offs, including the sample mean. Table 2 reports these estimates by CAM. Note that KR-21 refers to the value of $M(C)$ where $C = X_{pI}$. $\widehat{M(C_1)}$ refers to the lower cut-off, while $\widehat{M(C_2)}$ refers to the upper cut-off; i.e., that which separates Level 2 from Level 3.

Table 2

Error Components, Generalizability, and Dependability of Total Scores on APL Content Area Measures for Adult Education Students

Content Area Measure	$\hat{\sigma}^2(\delta)$	$\varepsilon\hat{\rho}^2$	$\hat{\sigma}^2(\Delta)$	KR-21	$\widehat{\underline{M}(\underline{C}_1)}$	$\widehat{\underline{M}(\underline{C}_2)}$
Community Resources	.00257	.94	.00304	.93	.97	.93
Occupational Knowledge	.00343	.92	.00388	.90	.95	.91
Consumer Economics	.00208	.94	.00271	.92	.96	.93
Health	.00349	.91	.00387	.90	.95	.91
Government and Law	.00364	.89	.00445	.87	.92	.91

As Table 2 shows, values of $\varepsilon\hat{\rho}^2$ are fairly high, ranging from .89 for Government and Law to .94 for Community Resources and Consumer Economics. Also, the values of $\widehat{\underline{M}(\underline{C}_2)}$. This reflects the fact that the sample means for each CAM were closer to the upper cutoff. In every case, the lower cutoff was set at 51% correct, and the upper cutoff was set at 76% correct. The sample means were 74% correct for Community Resources, 73% for Occupational Knowledge, 70% for Consumer Economics, 71% for Health, and 65% correct for the Government and Law CAM. In the case of Government and Law, values of $\widehat{\underline{M}(\underline{C})}$ differ by only .01. The mean score for the Government and Law CAM (65) falls close to halfway between 51% and 76%; thus, values of $(X_{pI} - C)^2$ are very similar for the two cutoffs.

The publishers of the APL Content Area Measures suggest that local decision makers may wish to modify cutoffs to suit local needs. Altering the cutoff, however, will result in a change in the dependability of the measures. The values listed in Table 2 under KR-21 represent the lowest possible values of $\widehat{\underline{M}(\underline{C})}$ for the data used in the development of the CAMs. It is also possible to set cutoffs in such a way as to increase the value of $\widehat{\underline{M}(\underline{C})}$. Figures 4 through 8 demonstrate the results of raising or lowering the value of C.

As can be seen in Figures 4 through 8, the generalizability coefficient $\varepsilon\hat{\rho}^2$ is totally unaffected by the value of the cutoff C. In other words, the position of the cutoff has no bearing on the ability of the test to rank order people. Note also, that the lowest value of $\widehat{\underline{M}(\underline{C})}$ is always below the $\varepsilon\hat{\rho}^2$ line. This is because the coefficient $\widehat{\underline{M}(\underline{C})}$ incorporates the variance due to item sampling in its definition of error, whereas $\varepsilon\hat{\rho}^2$ does not. Thus, by incorporating item variance in order to make absolute evaluations more meaningful, $\widehat{\underline{M}(\underline{C})}$ becomes a more conservative estimate of the precision of the test than $\varepsilon\hat{\rho}^2$.

Again, in reference to Figures 4 through 8, the values of $\widehat{\underline{M}(\underline{C})}$ increase rather slowly for Community Resources (Figure 4) and Consumer Economics (Figure 6) as C moves away from the sample mean. $\widehat{\underline{M}(\underline{C})}$ increases quite dramatically for Occupational Knowledge (Figure 5), Health (Figure 7) and Government and Law (Figure 8). These differences in slope reflect differences in the relative size of $\hat{\sigma}^2(\Delta)$ or error variance associated with each CAM. This is not to say that these three CAMs are inherently error prone but rather, that as the cutoff moves from the extremes to the mean, the dependability of the testing procedure declines more rapidly than it does in the Community Resources and Consumer Economics CAMs. In each CAM, the value of $\widehat{\underline{M}(\underline{C})}$ is nearly 1.0 when the cutoff is

COMMUNITY RESOURCES
CONTENT AREA MEASURE

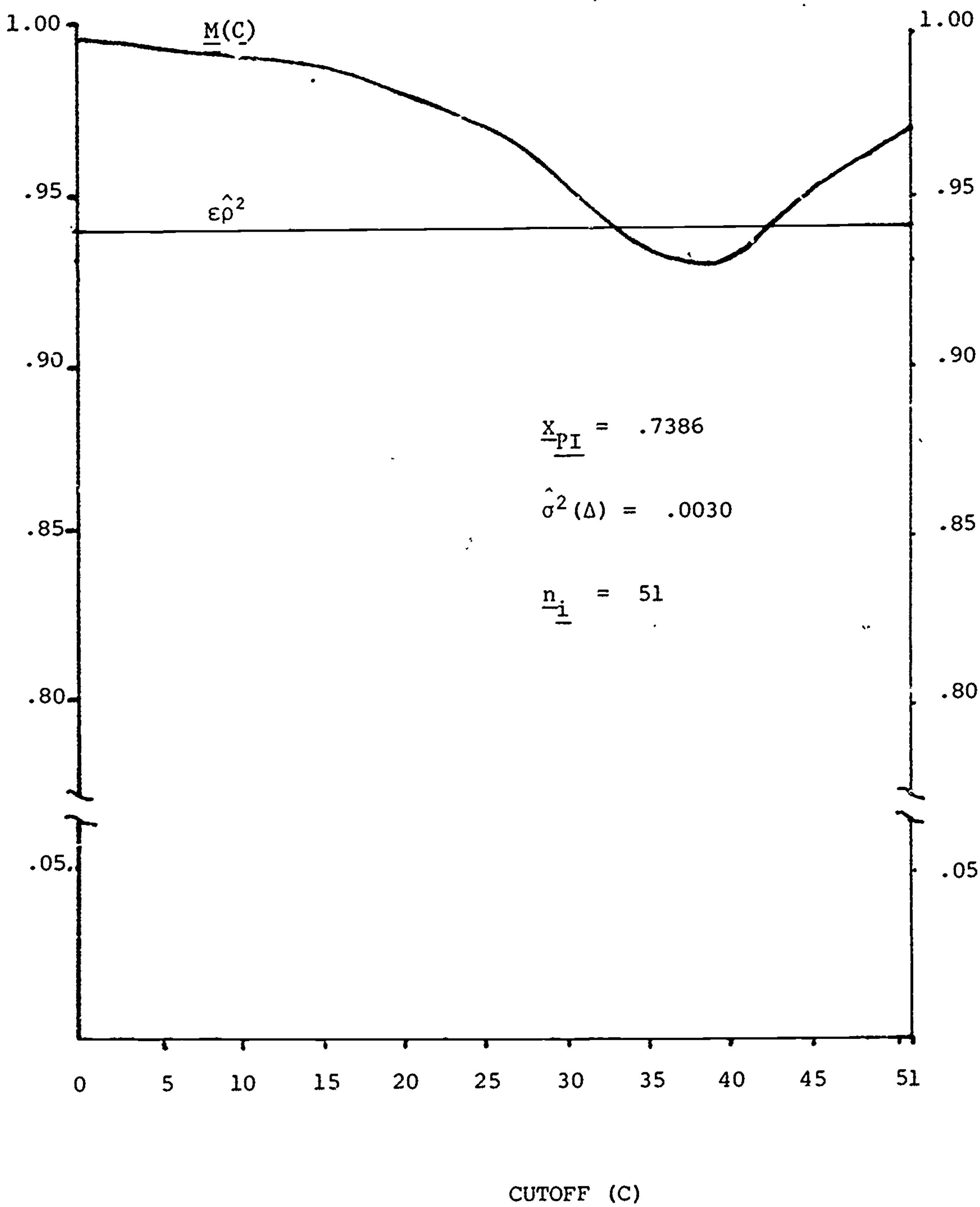

Figure 4. Generalizability/Dependability Coefficient as a Function of Cutoff.

OCCUPATIONAL KNOWLEDGE CONTENT AREA MEASURE

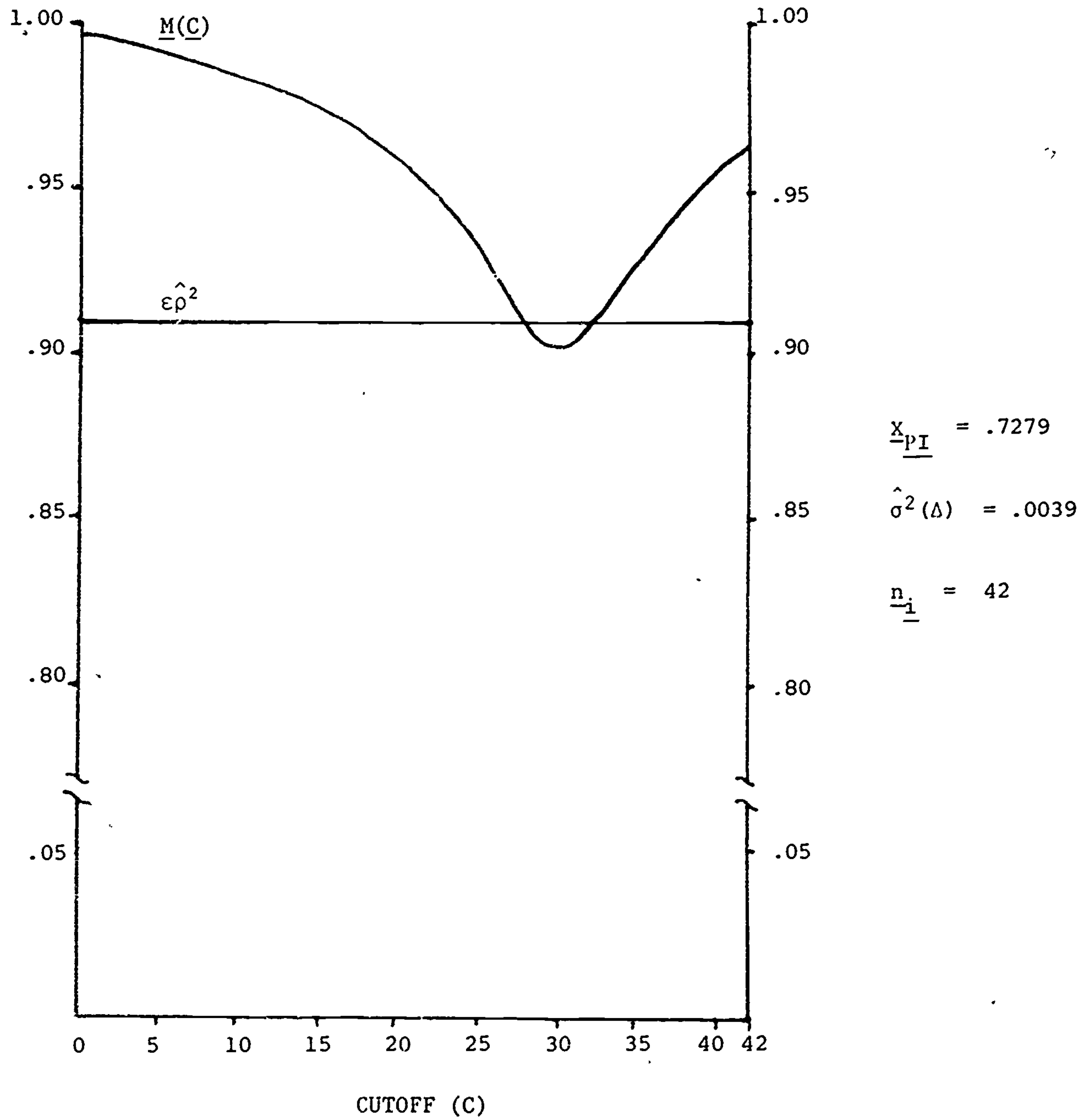

Figure 5. Generalizability/Dependability Coefficient as a Function of Cutoff

CONSUMER ECONOMICS
CONTENT AREA MEASURE

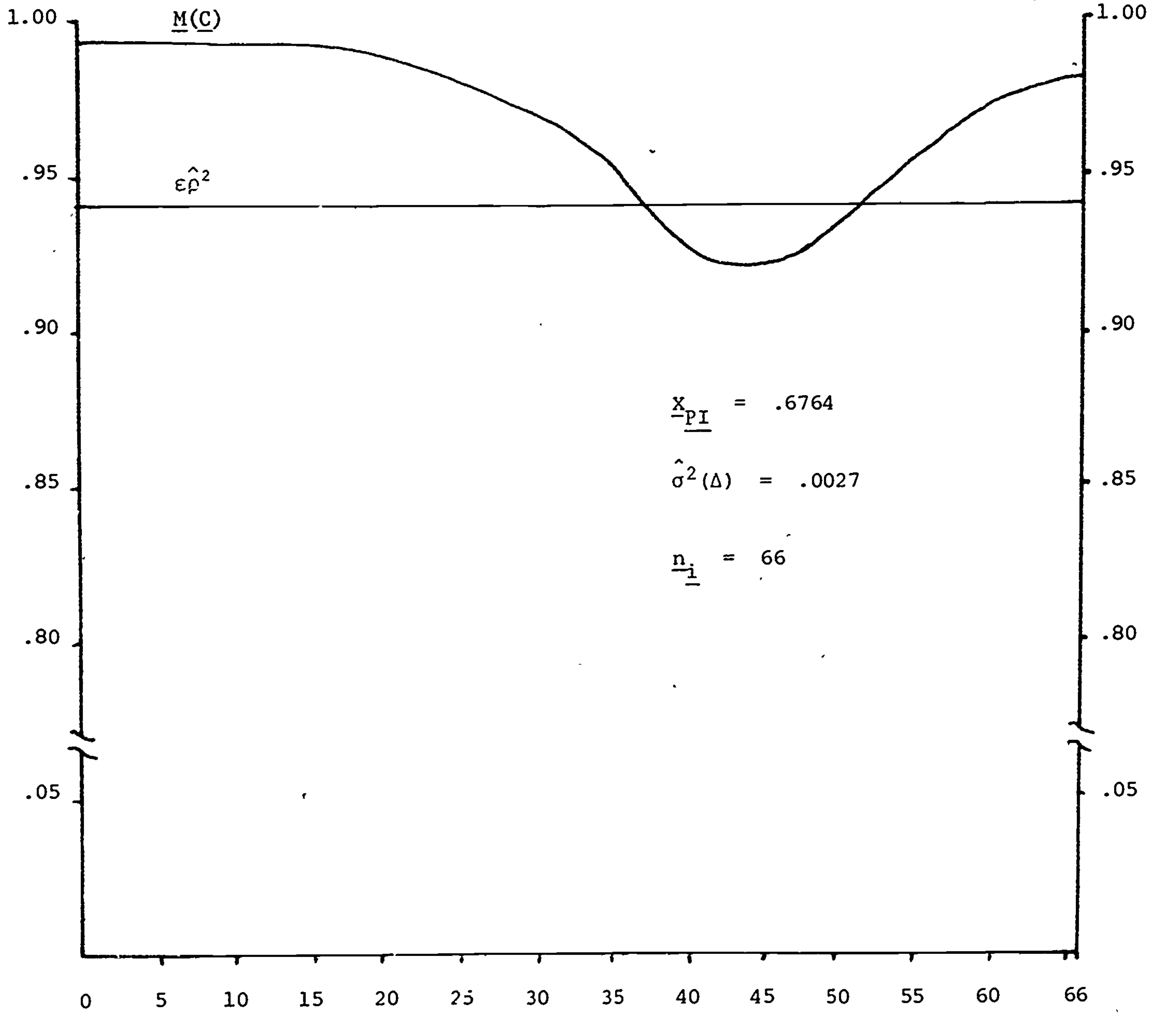

Figure 6. Generalizability/Dependability Coefficient as a Function of Cutoff.

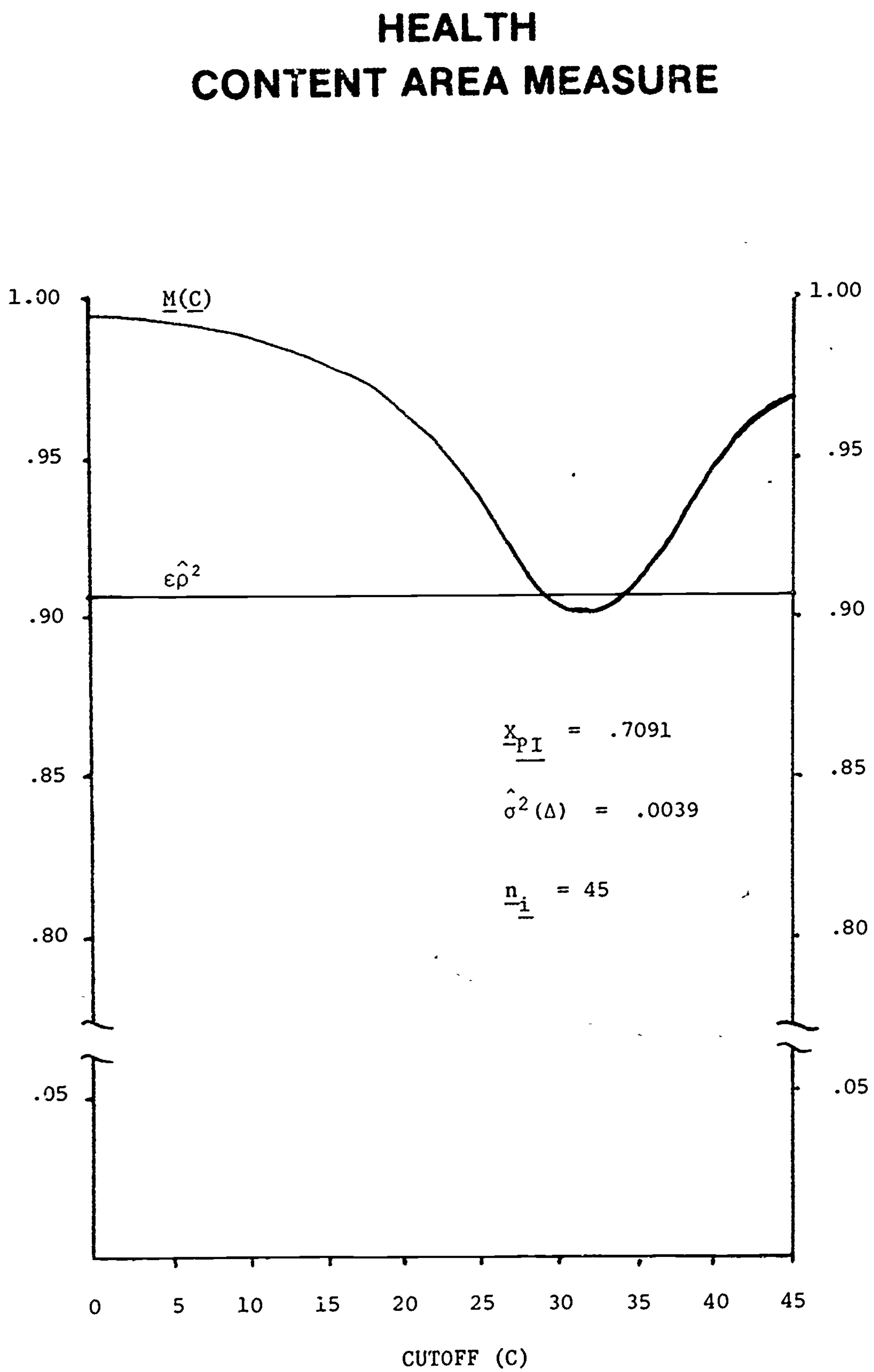

Figure 7. Generalizability/Dependability Coefficient as a Function of Cutoff.

GOVERNMENT AND LAW CONTENT AREA MEASURE

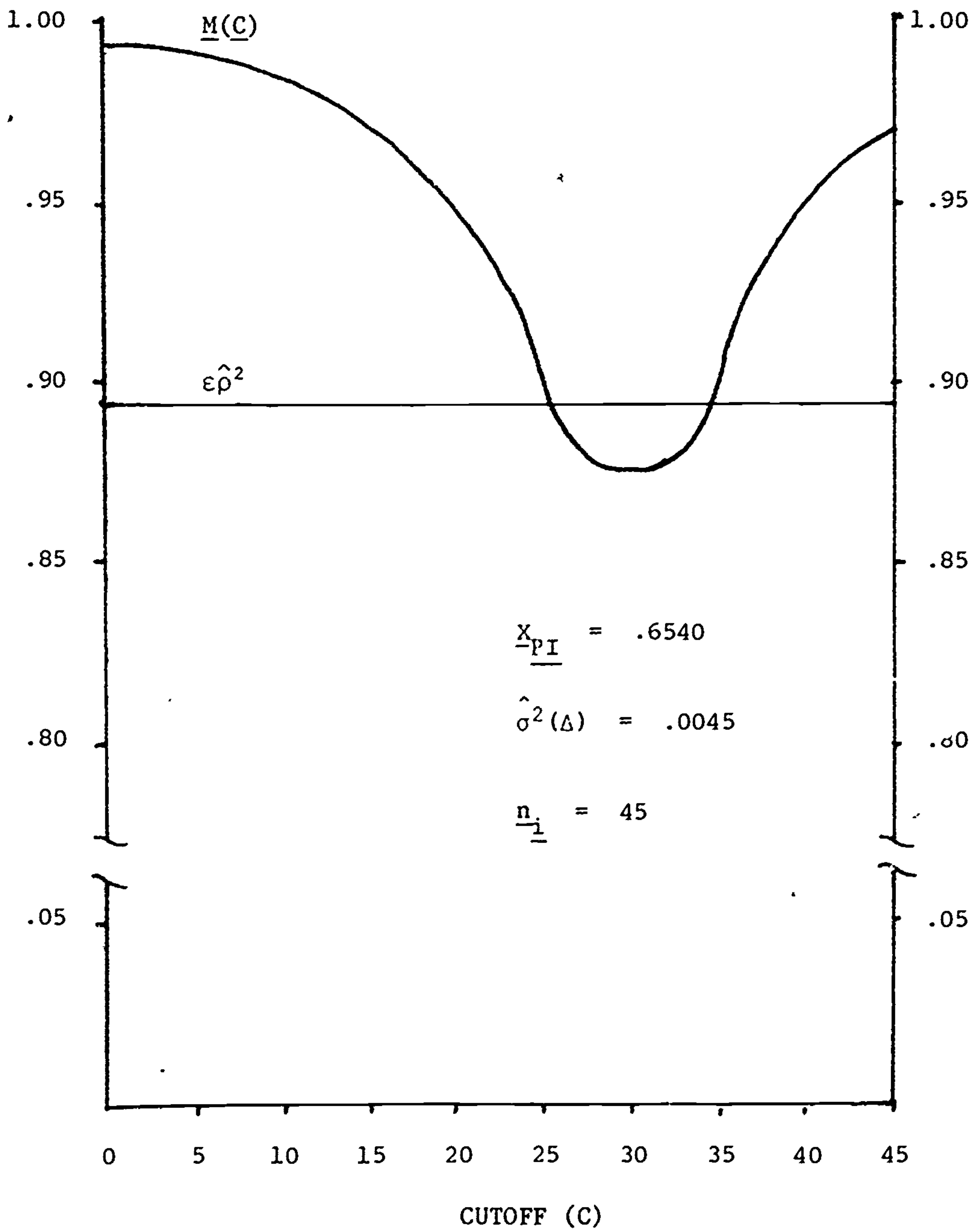

Figure 8. Generalizability/Dependability Coefficient as a Function of Cutoff.

set at 0 or 100% (trivial and highly unlikely cutoffs). Furthermore, respectable values of $\widehat{\underline{M}(\underline{C})}$ are maintained throughout the entire range of possible cutoff scores for each CAM.

Implications and Problems

Recalling now that the decision maker may raise or lower a cutoff as information justifies such action, one can see that there will be instances in which trade-offs between dependability and validity may become necessary. Assume for a moment that the cutoff score for Community Resources (Figure 4) that satisfied the conditions of equation (2) had been .74, or about 38 items correct. This would be the worst possible cutoff as far as dependability is concerned. Similar situations may arise if one uses Nedelsky's method, Ebel's, or any other content or criterion validity related method of setting cutoffs.

For the Community Resources CAM, the value of $\widehat{\underline{M}(\underline{C})}$ where C = 38 (74% correct) is .93. By either raising or lowering the cutoff, the decision maker could increase the dependability of the testing procedure. However, such action would also, in all likelihood, alter the probabilities of misclassification with respect to the external criterion.

In this particular case, the dilemma may not be very serious. The $\widehat{\underline{M}(\underline{C})}$ value of .93 is quite good. In other instances, it would be advisable for the decision maker to calculate or obtain values of KR-21 for the test to be used. If the value of KR-21 represents an acceptable level of $\widehat{\underline{M}(\underline{C})}$, than any value of C obtained through any cutoff setting procedure would be satisfactory.

Now suppose that for a given test the obtained value of KR-21 does not represent an acceptable level of dependability. This does not automatically mean that the test must be ruled out as an aid in making decisions about learners. Instead, this low value will limit the range of C. Should the value of C derived by equations (1) or (2) or any other procedure fall outside this restricted range, then adjustments are called for.

It might seem logical in such instances to ignore dependability indices and allow validity information alone to govern the setting of cutoffs. However, recall that a low value of $\widehat{\underline{M}(\underline{C})}$ (including KR-21) indicates a great deal of item variability relative to person variability. The model described in equation (2) does not allow for much item variability. Therefore, to the extent that item variability is large relative to person variability, the cutoff derived through equation (2) will be somewhat tenuous. For strictly content oriented models, item variability may also be a problem, depending on how narrowly one defined the domain of interest. The seriousness of this problem, given content oriented models, is not as obvious as in Emrick's (1971) model.

Another way to deal with the validity/dependability dilemma is to increase test length. Note in equation (17) that as n_i increases, $\widehat{\underline{M}(\underline{C})}$ will approach 1.0. If a value of C obtained through some procedure were to be inserted into equation (17) and a minimum acceptable value of $\widehat{M(C)}$ were set, then it would be possible to solve for n_i, the number of items needed to test at the desired cutoff and level of dependability. For locally produced tests this solution may be relatively easy to implement.

If, however, the decision maker is relying on standardized products, such a solution may be less appealing.

For tests such as the APL measures, where two or more cutoffs are suggested, a different kind of problem is possible It may turn out that data do not support a three group interpretation of test scores. In some instances, it may be more appropriate simply to classify learners into one of two categories, rather than into one of three or more categories. For example, adult education students who scored in the Average or Above Average range on some APL Survey (ACT, 1976) subtests may in some instructional settings be treated as similar to each other but collectively different from those who scored in the Below Average Range. A comparison of group score means would reveal whether or not such a strategy would be advisable. Cutoffs would then be adjusted accordingly.

Whatever the course taken in dealing with dependability/validity data, the crucial point is that somewhere in the process, the learner must derive some benefit over and above that which might be derived through random or arbitrary assignment. The benefit that will accrue to the learner will be a function of the correct classification of learner competencies and subsequent instruction. Within adult basic education, this focus on classification and instruction of individuals is seen as highly appropriate. Methods of assessing functional competency should be and generally are likewise individually oriented. Brennan and Kane (Brennan, 1977a, 1977b; Brennan & Kane, 1977, in press;

Kane & Brennan, 1977) have devised a frame of reference for expressing the dependability of such assessments. Examples drawn from the development of the APL Survey and Content Area Measures have been provided to demonstrate the usefulness of this frame of reference as well as of data obtained from non-test sources.

A systematic procedure has been described whereby the adult educator may make judgements not only about adult learners but about tests of functional competency as well. Definition, content validity, cirterion validity, and dependability as previously described all play important roles in the execution of this procedure.

NOTES

1. Conference on Strategies for Generating a National "Right to Read" Adult Movement, Raleigh, North Carolina, January, 1970.

2. Ronald M. Cervero, The Adult Performance Level Test: A measure of "functional competence"? Unpublished manuscript, University of Chicago, 1978.

3. This point is forcefully made by Gene Glass in "Standards and Criteria", a paper presented at the Seventh Annual Conference on Educational Assessment, 1977 and in "Postscript to 'Standards and criteria,'" a paper presented at the 1977-78 Winter conference on Measurement and Methodology of the Center for the Study of Evaluation, University of California - Los Angeles, January, 1978.

4. Jerry K. Williams, The APL: A minimal c[illegible]ency skills program. A presentation to the National Assessment of Educational Progress, Boulder, Colorado, June 14, 1977.

REFERENCES

American College Testing Program. User's guide: Adult APL survey. Iowa City, Iowa: Author, 1976

American College Testing Program. User's guide: Community resources content area measure. Iowa City, Iowa: Author, 1977 (a)

American College Testing Program. User's guide: Consumer economics content area measure. Iowa City, Iowa: Author, 1977.(b)

American College Testing Program. User's guide: Government and law content area measure. Iowa City, Iowa: Author, 1977.(c)

American College Testing Program. User's guide: Health content area measure. Iouwa City, Iowa: Author, 1977. (d)

American College Testing Program. User's guide: Occupational knowledge content area measure. Iowa City, Iowa: Author, 1977 (e)

American College Testing Program. Technical Supplement: APL content area measures. Iowa City, Iowa: Author, 1977 (f)

American Psychological Association. Standards for educational & psychological tests (rev. ed.). Washington, D.C.: American Psychological Association, 1974.

Brennan, R.L. Generalizability analyses: Principles and procedures. ACT Technical Bulletin No. 26. Iowa City, Iowa: American College Testing Program, September, 1977. (a)

Brennan, R.L., KR-21 and lower limits of an index of dependability for mastery tests. ACT Technical Bulletin No. 27. Iowa City, Iowa: American College Testing Program, December 1977. (b)

Brennan, R.L., & Kane, M.T. An index of dependability for mastery tests. Journal of Educational Measurement. 1977, 14, 277-289.

Brennan, R.L., & Kane, M.T. Signal/noise ratios for domain-referenced tests. Psychometrika, in press.

Cronbach, L.J., Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334

Cronbach, L.J., Gleser, G.C., Nanada, H., & Rajaratnam, R. The dependability of behavioral measurements. New York: Wiley, 1972.

Ebel, R.L. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

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

Griffith, W.S., & Cervero, R.M. The Adult Performance Level Program: A serious and deliberate examination. Adult Education, 1977, 27, 209-224.

Guilford, J.P. Psychometric methods (2nd edition) New York: McGraw-Hill, 1954.

Jencks, C., Smith, M., Acland, H., Bane, M.J., Cohen, D., Gintis, H. Heyns, B., & Michelson, S. Inequality: A reassessment of the effect of family and schooling in America. New York: Harper & Row, 1972.

Kane, M.T., & Brennan, R.L. Agreement coefficient as indices of dependability for domain-referenced tests. ACT Technical Bulletin No. 28. Iowa City, Iowa: The American College Testing Program, 1977.

Kuder, G.F., & Richardson, M.W. The theory of the estimation of test reliability. Psychometrika, 1937, 2, 151-160

Lord, F.M., & Novick, M.R. Statisitical theories of mental test scores. Reading, Massachusetts: Addison Wesley, 1968.

Meskauskas, J.A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard setting. Review of Educational Research, 1976, 46, 133-158.

Nafziger, D.H., Thompson, R.B., Hiscox, M.D., & Owen, T.R. Tests of functional adult literacy: An evaluation of currently available instruments. Portland, Oregon: Northwest Regional Educational Laboratory, 1976.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Northcutt, N. Functional literacy for adults: A status report for the Adult Performance Level study. Austin, Texas: The University of Texas at Austin 1974. (ERIC Document Reproduction Service No. ED 091 762).

Northcutt, N., Selz, N., Shelton, E., & Nyer, L. Adult functional competency: A summary. Austin, Texas: The University of Texas at Austin, 1975. (ERIC Document Reproduction Service No. ED 114 609).

Novick, M.R. , High school attainment: An example of a computer-assisted Bayesian approach to data analysis. International statistical Review, 1973, 41, 264-271.

Shoemaker, D.M. Principles and procedures of multiple matrix sampling. Cambridge: Ballinger, 1973.