DOCUMENT RESUME

ED 211 935 CS 006 442

AUTHOR Fisher, Donald
TITLE Assessment of Reading Competencies. Literacy: Meeting the Challenge.
INSTITUTION Office of Education (DHEW), Washington, D.C. Right to Read Program.
PUB DATE 80
NOTE 33p.; Paper presented at the National Right to Read Conference (Washington, DC, May 27-29, 1978). For related documents see ED 190 997 and CS 006 443-449.
AVAILABLE FROM Superintendent of Documents, Government Printing Office, Washington, DC 20402.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Literacy; Reading Achievement; *Reading Instruction; Reading Programs; *Reading Tests; *Standardized Tests; Test Construction; Testing Problems; *Test Validity

ABSTRACT

The first of eight related documents, this booklet is part of a series of papers presented at the 1978 National Right to Read Conference examining issues and problems in literacy. In its examination of standardized reading competency tests, the booklet first offers a definition of the kind of test it will consider and the criteria the test must satisfy to be deemed valid: content validity, construct validity, concurrent validity, and predictive validity. The paper reviews existing tests and offers approaches to achieving validity according to each criterion. It concludes that the education profession should not continue administering standardized tests in their present form. (HTH)

LITERACY: MEETING THE CHALLENGE

Assessment of Reading Competencies

Donald Eisher

The material in this booklet was prepared pursuant to a contract with the Right to Read Program, U.S. Office of Education, Department of Health, Education, and Welfare. Contractors undertaking such work are encouraged to express freely their professional judgments. The content does not necessarily reflect Office of Education policy or views.

The material in this booklet was presented at the National Right to Read Conference, Washington, D.C., May 27-29, 1978. The material was edited by the staff of the National Institute of Advanced Study which conducted the Conference under contract from the U.S. Office of Education.

FOREWORD

A major goal of the Right to Read Program has been to disseminate information about the status of literacy education, successful products, practices and current research finding in order to improve the instruction of reading. Over the years, a central vehicle for dissemination have been Right to Read conferences and seminars. In June 1978, approximately 350 Right to Read project directors and staff from State and local education and non profit agencies convened in Washington, D.C. to consider Literacy. Meeting the Challenge.

The conference focused on three major areas:

- examination of current literacy problems and issues
- assessment of accomplishments and potential resolutions regarding literacy issues; and
- exchange and dissemination of ideas and materials on successful practices toward increasing literacy in the United States

All levels of education, preschool through adult, were considered.

The response to the Conference was such that we have decided to publish the papers in a series of individual publications. Additional titles in the series are listed separately as well as directions for ordering copies.

SHIRLEY A. JACKSON
Director
Basic Skills Program

LITERACY MEETING THE CHALLENGE

A Series of Papers Presented at the
National Right to Read Conference
May 1978

SUMMARY

Overview

For the past fifteen years the validity of standardized tests, including those purporting to measure reading achievement, has been frequently called into question. Beginning with a definition of the kind of test it will consider and the criteria the test must satisfy to be deemed valid, this paper reviews existing tests in the light of each criterion in turn. It then offers approaches to achieving validity, again according to each criterion. It concludes that the profession should not continue administering standardized tests in their present form.

Definitions

Asserting that validity depends upon situation, the author defines as the test he will consider a general measure of children's functional literacy skills that will be used both for purposes of accountability and to identify minimal reading competencies. He then introduces the four criteria that test must satisfy: content validity, construct validity, concurrent validity, and predictive validity.

Threats to the Validity of Present Tests

No one type of instrument should bear the brunt of criticism because all four in present use fail to satisfy the criteria that define validity. The manner in which norm-referenced tests are constructed virtually precludes content validity. Objectives-referenced tests also fall short of content validity, first, because no substantial evidence links the sub-skills measured with the skills required for functional literacy, second, because literacy skills themselves have not been firmly established, and third, because the domain from which test items derive is undefined. The multiple-choice format common to standardized tests makes construct validity virtually unattainable because it does not allow students to give their reasons for choosing an answer and hence provides no assurance that their errors result from deficiencies in the skills that the questions intend to measure. The means for determining concurrent validity remain incomplete; in default of a reliable cut-off score, there is no assurance that any test identifies all and only the masters of the skills measured. The author bypasses the problems of determining predictive validity, first, because studies of existing tests rarely discover significant correlations between scores and performance in later life and second, because the tests studied were not constructed to predict what the test under consideration should predict: adult success in life.

Approaches to Improving Validity

Drawing upon approaches proposed and actually taken, the author offers methods of constructing functional literacy measures that would meet the objections set forth above. To assure that items are representative of the appropriate domain, i.e., to achieve content validity, it would be possible to adopt the approach taken by Army researchers, who ascertained what materials and for what purposes soldiers read in connection with their work and, from this information, defined job-related reading tasks. More practicably, it would be possible to ask content specialists to rate the relevance of items to the domain or to ascertain whether identically informed item writers could construct equivalent tests. Acknowledging that steps to improve construct validity, though many, are neither so simple nor so attractive, the author confines himself to arguing that students taking multiple-choice tests must have the chance to explain their answers. To arrive at a reliable cut-off score, a prerequisite for establishing concurrent validity, minimum passing levels for each item could be determined by experts and then totalled. The author refers to statistical techniques that might correct for measurement errors that result in misclassification. Finally, while acknowledging the naivete of trying to correlate any set of items to success or failure in life, the author presents two approaches to improving predictive validity, both illustrated by the Adult Performance Level Project.

Conclusion

Present standardized tests risk misclassifying students and hence not only fail to aid diagnosis and remediation, but perpetrate injustice. Since we cannot claim ignorance of the problem, we have a duty to confront it.

ASSESSMENT OF READING COMPETENCIES

Introduction

Ladies and gentlemen, it is indeed an honor and a pleasure to be here today. I hope to make our next hour as enjoyable as it is instructive. To this end, I have left out details which should perhaps have been included. Worse yet, I may have included too many tired or worn out issues. If I do not succeed for some of you, I trust you will understand it is not from a want of effort. As most of you are aware, the focus of this morning's talk will be on measures of reading achievement. In particular, we will ask whether the measures of reading achievement really identify what we claim they measure. And if the measures fail to identify what is claimed of them, then we will ask what can be done to improve them. So much for my prefatory remarks.

We as educators, we as parents, we as students, we all are no longer innocent. These are perhaps harsh words to begin a talk with. But I believe they are justified. A little over fifteen years ago a book was published which created some controversy. The book, aptly titled *The Tyranny of Testing* (Hoffman, 1962), was the beginning of an end to our innocence. Traditional standardized tests were vigorously criticized by the author of this book. The producers of standardized tests were quick to retaliate (see for example, "Explanation of Multiple-Choice Testing," 1961). The controversy raged for awhile, but then seemed to fade. Perhaps many hoped it were no more than a tempest in a teapot. However, matters heated up again in 1972. The National Education Association passed a resolution calling for a moratorium on standardized testing. The National Association for the Advancement of Colored People issued a similar statement in the spring of 1975. The Association for Supervision and Curriculum Development and American Association of School Administrators, while not calling directly for a moratorium, have used strong language about the need to reconsider uses of standardized tests (Perrone, 1977).

The debate has been pursued at length in many of the most respected journals in education. *Phi Delta Kappan* devoted this month's issue (May 1978) to the use of standardized tests (see articles by Brickell, 1978, Cawelti, 1978, Frener, 1978, Glass, 1978, Nathan and Jennings, 1978, Popham, 1978). *Today's Education* devoted their March/April issue to the problems surrounding standardized testing (see articles by Engel, 1977, McKenna, 1977, National Education Task Force on Testing, 1977, Taylor, 1977). The journal, *National Elementary Principal*, devoted two issues in 1975 to standardized testing. Yet, standardized tests are perhaps more in use today than they ever were before.

The litany could continue for days, perhaps weeks. Fortunately, we can stop here without doing damage to the point being made. Again, we are no longer innocent. We cannot easily claim to be ignorant of the existence of the controversy surrounding the use of standardized tests. We have been literally besieged with arguments both for and against the use of such measures. But, knowing that such arguments exist is one thing. Knowing which of these arguments to believe is quite another. I cannot hope to offer a complete treatment of the arguments bandied about. I do not pretend to be unbiased. But I can hope to present a few of the more salient issues as clearly and as forcefully as possible. Issues which bear repeating if already known, issues which deserve a hearing if unfamiliar. So, we now turn to a discussion of the major criticisms levelled against the use of standardized tests. In particular, we will focus on those criticisms which bear in one way or another on the purported validity of standardized tests of reading.

Accountability, Validity and Minimal Competencies

At least four criticisms have been raised against standardized tests of reading achievement. The content validity of the tests, the construct validity, the predictive validity and the concurrent validity have all received their share of criticism. In order to expand on these criticisms we need a more detailed description of the villain and a more detailed description of what we hope to accomplish with our measure of reading achievement. The villains are standardized tests, but certainly not all standardized tests or all aspects of any one standardized test need be villainous. The goal of our testing is a valid measure of reading competencies, but certainly some competencies are important in one situation, and not important in others. These points cannot be made too strongly. A test is valid only in a certain context or situation. So, a test valid in one situation may not be valid in another situation.(Anastasi, 1965). The criteria for validity may themselves conceivably change from situation to situation.

Thus, it is only proper to ask me at this point just what sort of situation I have in mind. First, I am assuming that one is looking for a general measure of children's functional literacy skills. Second, I am assuming that the measure is to be used for purposes of accountability. And third, I am assuming that the measure will be used to identify what have been called minimal reading competencies. Interest in functional literacy, accountability and minimal competencies have more or less gone hand in hand. The interest is centered at all levels of government, national, State, and local. (Fisher, 1978) As of March 15 of this year, 33 States had taken some action to mandate the setting of minimum competency standards for elementary and secondary students. (Pipho, 1978) And many of these states are using measures of functional literacy as standards of accountability. Typical functional literacy items include those used on a test given by the Educational Testing Service (see Figures 1 through 5). The stem for the items are, in order:

1. Place a circle around the bottle of liquid that would be safe to drink.
2. Look at the train schedules. Put a circle around the time the daily train leaving Trenton at 4:46 p.m. arrives in Washington.
3. Put a circle around the label that would be the best one to put on a box used to mail something easily broken.
4. Look at the garment tags. Circle the two tags that indicate the garments are made from 100 percent Polyester.
5. Look at the application for employment. Put an X in the space where you would write the names and addresses of someone to notify in case of an emergency.

With the content and purpose of the measure in hand, we can go on to delineate the criteria which such a measure must satisfy in order to be considered valid. First, the criterion of representativeness or content validity must be satisfied. The materials on the test should be representative of the materials it is thought important for the student to read. And the questions asked on the test should be representative of the sorts of questions it is thought important for the student to answer. (See Bormuth, 1973, for a much more complete exposition.) Second, the criterion of fairness or construct validity should be satisfied. Children should not be penalized for a defensible answer, even though this answer deviates from the one originally thought correct. The existence of defensible answers is much more ubiquitous than one might at first imagine if one can believe what one reads. There are other aspects to the criterion of fairness. These aspects need not be mentioned until later. Third, the criterion of present relevance or concurrent validity must be satisfied. In the present case this means that the measure must be able to differentiate between masters and nonmasters of functional literacy skills. And fourth, the criterion of future relevance or predictive validity must be satisfied. In the present case this means that later in life the masters must possess the minimal competencies needed to weave their way through the warp and woof of daily existence. Conversely, the nonmasters must not have the skills needed to function at some minimum level in society. Type I measures will be used as a shorthand to refer to instruments which meet these standards of validity.

The situation has been well-established. That is, the measure of reading achievement is to be used for the purpose of accountability. It remains to identify the villain.

Four Criticisms of the Villain

The villain is not any one sort of instrument. Criterion-referenced measures of reading achievement objectives-referenced measures of reading achievement all can be improved upon. If there is one message which you take home with you at the end of this hour or so it is just this. Labels alone do not make a

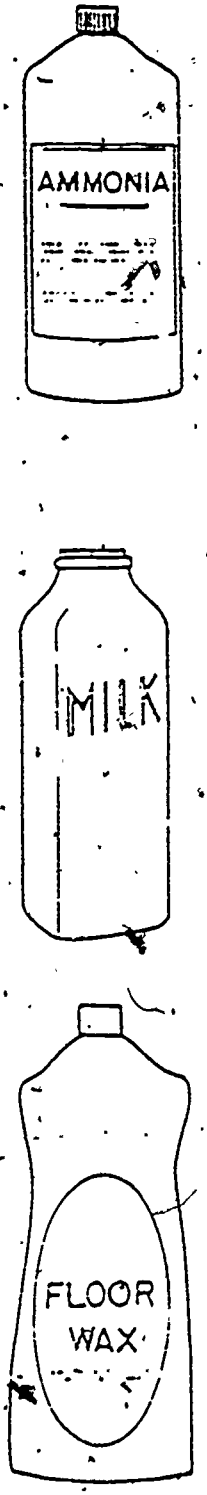
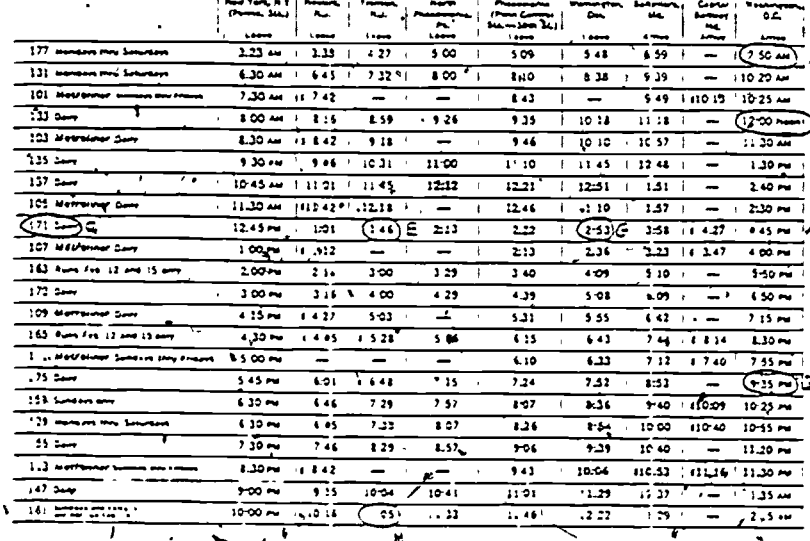


FIGURE 1

Book 4: Item 2

New York – Washington



Washington – New York

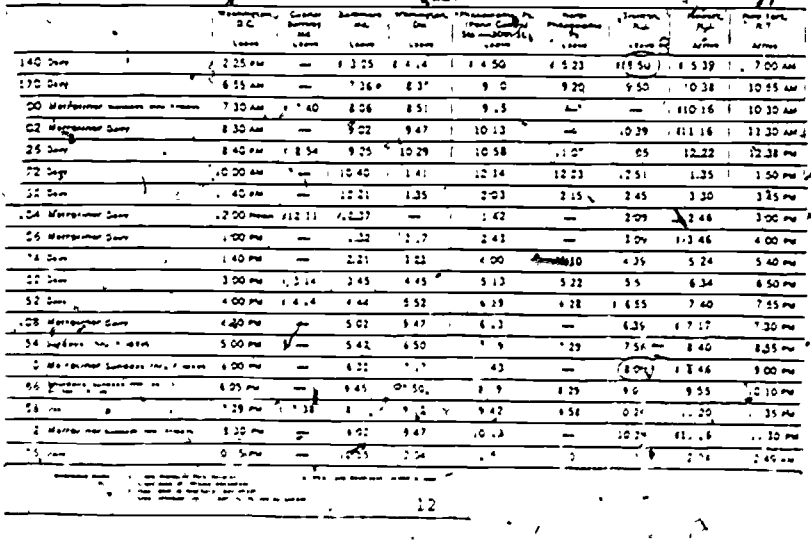


FIGURE 2
Book 1 Item 13

measure valid. This message is simple enough, but it too often gets lost in the rhetoric of the testing controversy. A test must be measured against the four types of validity we set out. And how do present day measures stack up against these criteria? An attempt to answer this question follows immediately. We will look separately at content validity, construct validity, concurrent validity and predictive validity.

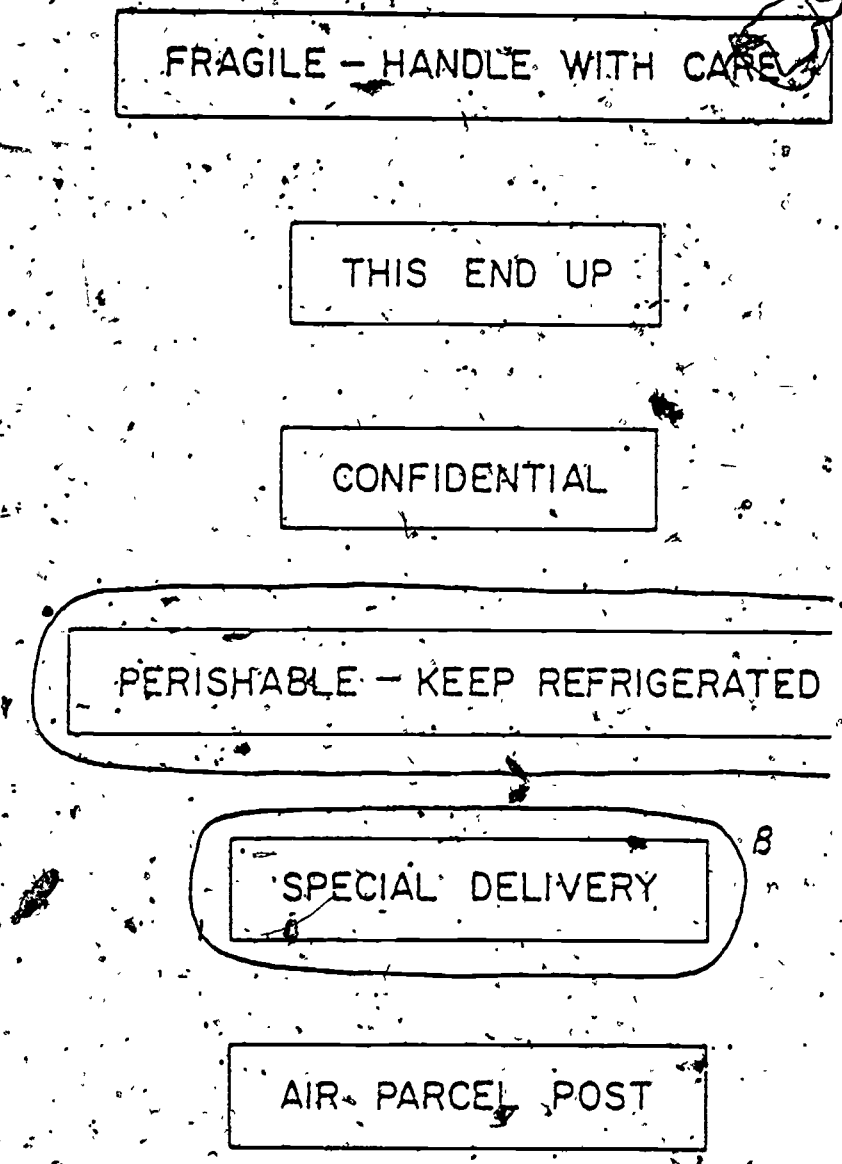


FIGURE 3
Book 3: Item 1

Content Validity

I spoke earlier of the criterion of representativeness or content validity. It will be argued here that both norm-referenced measures and objectives-referenced measures pose threats to the content validity of type I tests. First, consider norm-referenced tests. We can, in fact, say something very strong

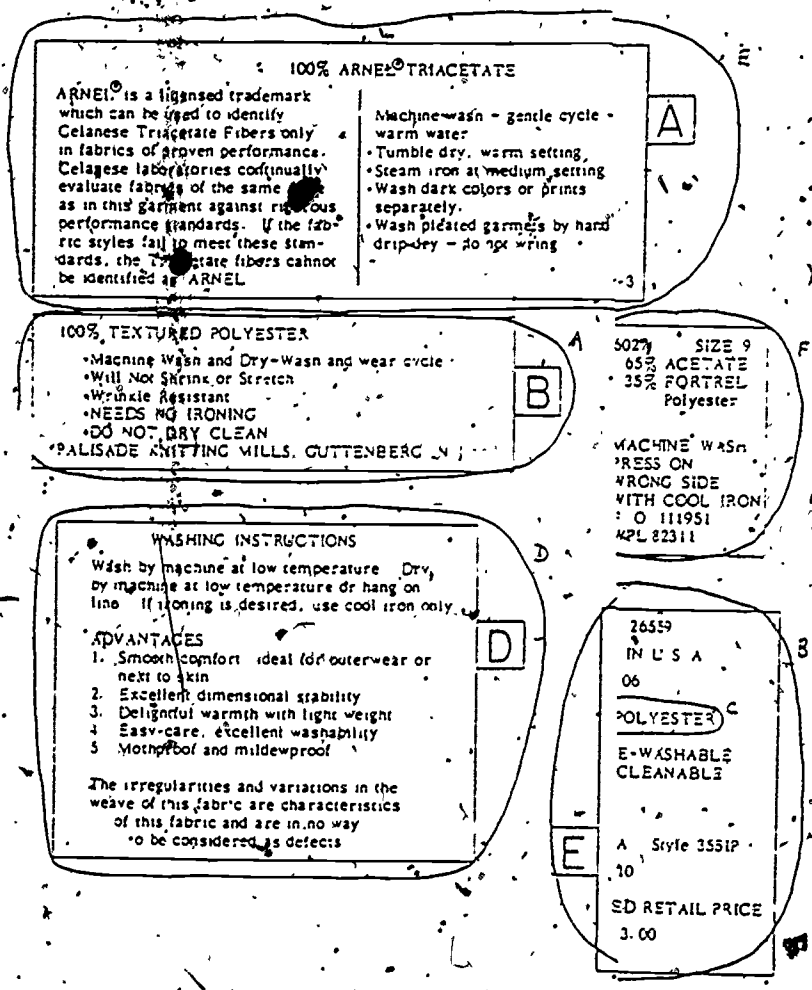
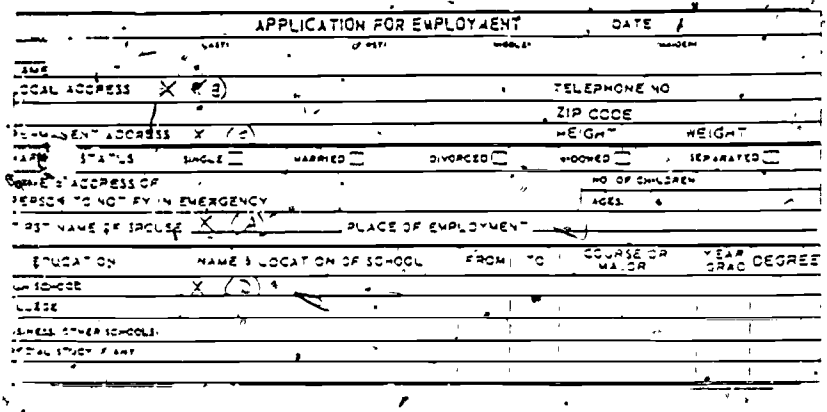


FIGURE 4
Book 4: Item 9

about the content validity of norm-referenced measures. We can say that in principle norm-referenced measures are designed to identify differences in ability or achievement between students. Therefore, items will be excluded from the instrument which fail to differentiate between students. In the extreme case, an item may be answered correctly by everyone. This item would inevitably be jettisoned (Popham, 1974). If everyone answers the item correctly, then there is no reason to include it on a norm-referenced measure. In general, it is clear that such a procedure will not lead to a representative sample of behaviors or items. Items are systematically excluded and included on the basis of their difficulty, not on the basis of their representativeness.



APPLICATION FOR EMPLOYMENT DATE

NAME

LOCAL ADDRESS TELEPHONE NO

ZIP CODE

PERMANENT ADDRESS HEIGHT WEIGHT

MARITAL STATUS SINGLE MARRIED DIVORCED WIDOWED SEPARATED

NAME & ADDRESS OF NO OF CHILDREN

PERSON TO NOTIFY IN EMERGENCY AGES

FIRST NAME OF SPOUSE PLACE OF EMPLOYMENT

EDUCATION	NAME & LOCATION OF SCHOOL	FROM	TO	COURSE OR MAJOR	YEAR GRAD	DEGREE
HIGH SCHOOL						
COLLEGE						
GENERAL OTHER SCHOOLS						
SPECIAL STUDY OR ART						

FIGURE 5
Book 5: Item 5

Some members of the audience may accept the technical point, yet find it rather unexciting, lacking any real oomph. Hopefully an example can give life to the point. Figure 6 contains three sample items which might appear on a pilot test of reading achievement in the 6th grade. Suppose one finds that over 99 percent of the students answer the first item correctly, roughly 50 percent of the students answer the second item correctly, and fewer than one percent of the students answer the third item correctly. In general, one would not expect to find the first and third items appearing on the final version of the test. They would be excluded because they fail to differentiate between students. Yet they are clearly important items. By systematically excluding very easy and very difficult items one may well be excluding those items, those reading behaviors, which one would most like to see measured.

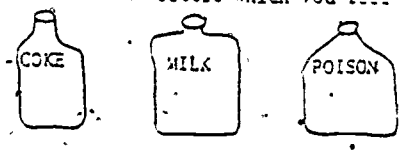
It has been argued that the content validity of type I measures is threatened by the manner in which norm-referenced tests are constructed. It is also the case that the content validity of type I measures can be defined as those tests which specify in great detail all the reading competencies or behaviors the individual is thought to need. Some tests have specified upwards of 350 skills. Global tasks are dissected, placed under the microscope, and claimed to consist of so many minute behaviors. In principle, such an endeavor is laudable enough. In practice, such efforts turn out to be rather dreary. This is so for at least two reasons.

First, there is little, if any, evidence which links the specific sub-skills measured on objectives-referenced tests with the reading skills one might need as a functional literate. Indeed, this is not surprising. Functional literacy is very much a part of literacy. And the specific skills needed to be literate, to understand what one has read, have never been well-defined. (Farr, 1969) Thus, the set of skills selected for testing on objectives-referenced measures may well not be representative of the larger set of skills needed for functional literacy. This point of view receives further support when one goes so far as to

analyze the errors made on tests of comprehension. Particularly instructive is the analysis of errors in the functional literacy measure administered by the Educational Testing Service. (Murphy, 1975) There were two phases to this analysis. In the first phase, the answer booklets from a previous administration of the functional literacy test were examined. Fully 85 percent of the incorrect responses could not be categorized. If the relation between reading competency sub-skills and general reading behaviors were transparent, one would expect to find the classification task easy enough under most circumstances. One could say, "This person made an error here because they didn't have such and such a sub-skill," and so on. Now if one can't relate putative sub-skills to actual reading tasks or demands, one has the beginnings of a problem. For the possibility is raised that some of the objectives are truly irrelevant to the demands written materials place upon the reader. Thus, the set of skills selected for testing on objectives-referenced measures may well not be representative of the large set of skills needed for functional literacy.

The second phase of the analysis makes the same point still more forcefully. Examinees were asked to elaborate on their answers as they went along. Other than vocabulary and item format, the answers on the whole were not

1. Draw a circle around the bottle which you feel is unsafe to drink from.



2. Dime is to nickel as four is to ____

(a) 11

(b) 1

(c) 5

(d) 2

3. Which of these containers lying by the side of the road would you be most apt to report to the authorities?

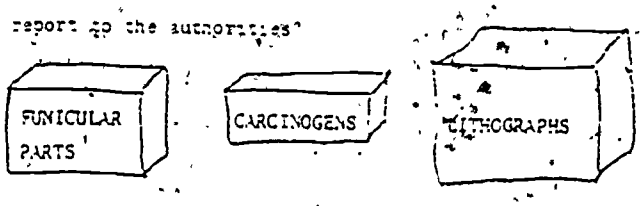


FIGURE 6

amenable to any rigid sort of categorization or explanation. In fact the responses to a particular item were often unique to one individual respondent.

Several examples of the rather unique way in which individuals respond are offered by Murphy (1975). They bear repeating here:

1. In a list in which the respondent is to choose an entry corresponding to baby's clothes, the entry *hampers* appeared. A respondent who chose that entry explained that he thought *hampers* might be like "Pampers" - a commercial product of disposable baby's diapers.
2. A list contained several amounts of alcohol and the effects associated with drinking such amounts. A respondent was asked to circle the amount associated with a given effect. He circled a greater amount and gave as his reason his disagreement with the chart. He judged that the effect would be associated with a greater amount of alcohol.
3. A doctor's bill listed the amount owed. A respondent circled a higher amount listed on the bill because it corresponded more closely to her own latest doctor's bill.

Do these responses appear to be due to the absence of any of the sub-skills one most often sees on tests of reading comprehension? I think the answer would have to be no. Objectives-referenced tests at the moment simply do not possess the content validity required of a type I measure. It should be mentioned that the violation of content validity being considered here could also be construed as a violation of construct validity.

The content validity of objectives-referenced measures is threatened for still another reason. Ultimately, we want to be able to generalize from the type I measure to the domain under consideration. Consider a behavioral objective such as, "The student shall be able to spell 80 percent of the list of 50 words presented to him in a period of less than ten minutes." Such an objective says nothing about the domain of words from which the test items are selected. As such we cannot immediately generalize to some larger set of words. Our knowledge about the student's performance is confined to the list of words on the test. When the domain of behaviors is not clearly established, one has what Popham (1974) calls a 'cloud-referenced test.'

Construct Validity

Three related criticisms have been levelled against the construct validity of various of the standardized tests. It will be remembered that the construct validity of an instrument is threatened when an item purports to measure one cognitive activity but actually measures another. The first of these related criticisms is aimed specifically at the multiple-choice format of most standardized tests. The critics find all manner of things wrong with the

multiple-choice format (see for example Hoffman, 1962). We will concentrate on only one. In particular, I will focus on that criticism which faults the multiple-choice format for not allowing a student to indicate his or her reasons for choosing a particular item. As such, one never knows why a student answered the item correctly, or for that matter incorrectly. In such a case one could say that the measure is hierarchically opaque. If answers were unambiguously right or wrong, there would be one less reason to argue against the multiple-choice format. But answers are not always clear cut.

A particularly invidious sort of item is the verbal analogy. We have already seen an item of this type. (Figure 6, item #2) Note that one can easily provide reasons which lead to the choice of any one of the alternatives. Choice (d) is correct if one argues that the value of a nickel is one-half the value of a dime, just as two is one-half of four. Choices (a), (b), and (c) are correct if one decides that ten, the dime, is to five, the nickel, as even is to odd. Choice (c) is correct if one decides that just as a nickel is the least smallest coin less than a dime, so there is a least smallest integer less than four, namely three. Choice (b) is correct if one decides a dime is 5 more units than a penny just as 4 is five more units than 1. And choice (a) is correct if one decides that a dime, which has four letters, is to nickel, which has six letters, just as four, which has four letters, is to eleven, which has six letters.

Note that this is a problem intrinsic to the testing of verbal analogies with the multiple-choice format. One can almost always (I hesitate to say always) find reasons for choosing one alternative over another. And in general the multiple-choice format is to be avoided. One simply cannot know whether students are identifying a 'correct' answer for the wrong reasons, or identifying an 'incorrect' answer for the right reasons. The threat to construct validity is potentially enormous when one fails to identify the reasons a person has chosen a particular alternative.

A second related criticism of construct validity again focuses on the hierarchical opaqueness of such measures. Remember that the measures are being called opaque because it is simply not possible to determine why a student has chosen a particular alternative. Now suppose one wanted to know whether an English-speaking person could broad jump four feet. One might measure a four foot span, putting markers at the beginning and end of the span. One would then go out and accost the first friendly neighbor run into. But instead of asking him or her to jump the four feet in English, suppose one asked him or her to jump the four feet in Chinese. Well it is quite clear that the person who speaks only English will not understand the request. It is equally clear that the person may well be able to leap the full four feet. The moral of the story is apparent enough. If one wants to test a person's broad jumping skills one doesn't at the same time test his language comprehension skills.

But what relevance does this have to measures of reading achievement, and in particular to the issue of construct validity? Figure 7 may contain ordinary

enough phrases for the members of this audience. But for one group of examinees the list of words was far from ordinary. The list of words were as foreign to them as Chinese is to a person that speaks only English. All the words appeared as part of the item stem on a test administered by the Educational Testing Service (Murphy, 1975) The test was a measure of functional literacy skills It was determined in later analyses of other individuals that these phrases were responsible for many errors. Unfortunately, these error analyses were done after 10,000 individuals had already been tested The test was not supposed to measure an examinee's understanding of the stem Yet because the measure was hierarchically opaque it did indeed do so

One no sooner leashes this argument on one's opponents, than one's opponents pipe up with what appears to be a cogent counterargument They claim that the fault lies not with the multiple-choice format but with the failure of the test makers. The counterargument calls for a proleptic defense of sorts. While I grant the opponent's counterargument in principle, I see no reason to admit defeat in fact Remember the conclusions of the error analysis done by the Education Testing Service reported in the previous section. The 'errors' so to speak were unique to each respondent Many of the alternatives might have been considered correct if the examinees had been allowed to voice the reason for their choices Since the reasons which lead to the particular choice of an alternative are so unique, it seems most unlikely that one can design a valid

to call up	ingredients	to operate
transplant	liver	classified
apparel	firearms	fuel
commencement	locker	stance
lives	extinguisher	minimum
toll	ingestion	severe
injection	correspondent	mild
series	to fill in	whom
creed	come together	experience
misstatement of fact	permanent	
confronts	pesticide	
recipe	circle	
	fourth	

FIGURE 7

multiple-choice test, a multiple-choice test which leaves no room for the idiosyncratic answer.

This same criticism extends to other situations. We say we are measuring a behavior such as determining the main idea of some prose passage. But can we infer that individuals who did not get the main idea are deficient in "main idea" skills? Probably not. We are more likely measuring something like vocabulary. We say we are measuring a behavior which has something to do with drawing inferences. But again, can we infer that the person who did not circle the correct answer is deficient in referencing skills? Again, probably not. We are just as apt to be measuring general familiarity with the test material. Some individuals might choose to argue at this point that the notion of abstract skills such as "inferencing" or "getting the main idea" are empty notions by themselves. For example, one does not draw inferences in a vacuum. One draws inferences with respect to a particular content or written passage. Therefore, it is legitimate to consider vocabulary as a component or inferencing skills. I know of no hard and fast argument against such a position. But I do know that such a position can only muddy the waters. For if I want everyone to do poorly on some test of inference, I can simply make the passage abstruse enough. In short, I can easily make it difficult if not impossible to know whether students do indeed possess anything like inferencing skills.

The third and final criticism is again focussed on the hierarchical opaqueness of the standardized measures. However, this time their potential for cultural bias is at issue. The existence of cultural bias in tests affects their construct validity in the same way the existence of obscure vocabulary affects the construct validity. Some studies have shown little change in the performance of students when the test is rewritten in the dialect favored by the students. The study I am going to report does find a change, a very large change. The study (Thurmond, 1977) was an attempt to measure the effect of black dialect on reading test performance of black and white high school students. Forty-six low achieving ninth-grade students were administered a standard English form and black dialect form of the reading subtest of the Stanford Diagnostic Reading Test. The dialect form was written so that the written language of the test approximated the exact oral sentence pattern of the black students taking the test. Results showed that black students administered the dialect form did significantly better (.05) than black students administered the standard English form. White students did significantly better than black students on the standard English form of the test. The means are reported in Figure 8. The results are especially striking when one realizes that as little as a point can make a full grade difference in reading level.

Concurrent Validity

There are three common ways of measuring the concurrent validity of norm-referenced tests. The test can be validated against some other already

established test. The measure can be compared with some other criterion, such as judges' rating of performance. Or the test can be validated using what is called the method of contrasted groups. For example, the test may be given to one group that is thought to possess the skills in question, and to another group that is thought not to possess the skills in question. These methods are appropriate to type I measures, but incomplete. The problem comes in finding a reasonable cut-off score, something that has not been attempted until recently on any large scale. The finding of a reasonable cut-off score is a problem for concurrent validity in the following sense. If the cut-off score is set too high, it will fail to identify all the masters of a given domain. Therefore, it will have less than perfect concurrent validity. Similarly, if the cut-off score is set too low, it will identify some nonmasters and masters. But I am getting ahead of myself. The new techniques for setting cut-off scores are best discussed in a later section.

Predictive Validity

At the moment we can also bypass the problems involved in the determination of predictive validity. This can be done for two reasons. First, consider extant measures where studies of predictive validity have been undertaken. In the numerous studies where test scores have been used to predict adult success one rarely if ever finds significant correlations (Nathan and Jennings, 1978). For example, the dissenting opinion of California

Test Form/Race	Black	White
dialect	30.3	30.7
standard	22.2	31.4

FIGURE 8

Supreme Court Justice Arthur Tobriner in the celebrated Bakke case cited numerous studies showing no correlation between high medical school admission test scores and quality performance as a physician later in life ("Regents of the University of California v. Allan Bakke").

> Indeed, the medical school's decision to deemphasize MCAT scores and grade-point averages for minorities is especially reasonable and invulnerable to constitutional challenge in light of numerous studies which reveal that, among qualified applicants, such academic credentials bear no significant correlation to an individual's eventual achievement in the medical profession. The findings of these studies are not surprising when one considers all of the nonacademic qualities—energy, compassion, empathy, dedication, dexterity, and the like—which make for a "successful" physician.

One more example is worth citing. A recently published *Phi Delta Kappan* article (Jennings and Nathan, 1977) cited research on the complete lack of correlation between high scores on the two major college admissions tests and success in adult life. The two tests considered were the American College Test and the College Entrance Examination Board.

There is yet a second reason we can bypass a critique of the predictive validity of previous measures. In general, such measures have not been constructed with regard to the sort of predictive validity we have in mind. Until recently no tests that I know of have been produced specifically for the purpose of predicting adult success in life. So, we cannot criticize earlier measures simply because few are around to criticize.

Summary

To summarize, we have examined the various threats to the validity of type I measures posed by traditional, standardized tests. It is not the name of the test so much as the manner of construction and the item format which identifies a measure as particularly offensive. The content validity and construct validity of some present measures was found wanting. The need for new approaches to concurrent and predictive validity was noted. Appropriately enough, it is now time to consider just such new approaches.

Recommended Type I Measures

Criticism is abundant. Constructive criticism is a bit more dear. And creative, plausible alternative suggestions and solutions are hardest to come by. Fortunately, this is one of those infrequent and happy occasions when alternatives are available and cheap. Of course, the relative abundance of alternatives to established ways does not preclude criticism. But at least it leaves the road to constructive action paved with possibilities. As in the last section, my remarks will all fall under the general rubric of validity. First then we turn to a discussion of the ways in which one might go about improving the validity of type I measures.

Content Validity

Norm-referenced and objectives-referenced tests were seen as threats to the content validity of a type I measure. In one way or another, the norm-referenced and objectives-referenced tests biased the content of these measures. In general, the questions and materials on these tests are not representative of the universe of written questions and materials. In the abstract, the steps one must take to help guarantee representativeness are clear enough (Ebel, 1962, Hively, Patterson, and Page, 1968) The steps are summed up by Hambleton et al (1978):

1. The domain must be specified clearly enough so that all items which could be written from the content domain to be tested must be written or known in advance of the final item selection process.
3. A random or stratified random sampling procedure must be used in the item selection process.

While these goals remain models to strive for, they are rarely if ever approached in practice So we need ways of approximating these goals. Two such approaches are discussed.

The first approach might be referred to as a hands-on attempt to retain content validity. Something akin to the research done by Dr. Stitcht, at present an Associate Director of the National Institute of Education, seems desirable (see Vineber et al , 1971) A brief review of this work is in order. In 1966, the United States Army initiated Project 100,000. Up to 100,000 individuals were to be let into the armed services who would previously have failed to gain entrance for reasons of health or measured intellect. The military needed to know how much, if any, literacy training these men would require. A sample of the Project 100,000 men suggested that much work lay ahead. Approximately 30 percent of the sample read below the fourth-grade level while almost 70 percent read below the sixth grade level By themselves these figures mean little The Army did not know the reading demands of the various occupational specialties it could expect the Project 100,000 men to enter. Nor did the Army know whether the scores on standardized tests of reading achievement were good predictors of job performance. Therefore, the Army sought to obtain information concerning the literacy demands of military jobs and the predictive power of reading and other related tests.

In order to determine the actual reading demands of personnel in Army jobs, research psychologists interviewed men at work. The men indicated both what they read and why they read it. The most frequently cited materials and tasks were eventually included on what came to be called job-related reading tasks. The face validity of such an approach is unimpeachable. Unfortunately, it exists in practice much less frequently than one might suppose.

The second approach is more likely to be the one adopted by the majority of individuals involved in the construction of functional literacy measures. Instead of actually sampling the domain of behaviors, judges are asked to indicate the relevance of an item to a particular domain. Any one of a number of schemes have been suggested. Rovinelli and Hambleton (1977) suggest three procedures, any one of which could be easily constructed. For example, they suggest that content specialists rate the relevance of an item to the domain being tested. One computes the mean of the ratings across content specialists for each item. Presumably one can then agree upon some cut-off score below which items are no longer considered appropriate to the measures of a given domain One can easily compute the variance associated with each item. This gives a measure of the agreement among content specialists.

Cronbach (1971) suggests what might be called a duplication experiment. A group of item writers is selected and randomly divided in half. They receive the same information about the domain to be tested. If the domain specifications are clear, and the sampling representative, then the tests should be equivalent. Clearly, these are only stop gap measures. However, since the potential harm of such methods seems minimal, and since the methods may indeed point up weak spots, they may be worth pursuing. It should be noted in passing that the beginnings of much more technically precise ways of sampling from a domain have appeared in the literature recently (see for example Bormuth, 1973; Hively, Patterson and Page, 1968). However, these methods are not yet applicable in any area quite as diffuse as functional literacy.

Construct Validity

The steps needed to improve the construct validity of type I measures are neither as simple nor initially as appealing as the steps required to improve the content validity of such measures. Many approaches are possible. I will present only one. It seems to me imperative that the student be given a chance to explain his choice of a particular alternative while a multiple-choice test is in progress. And furthermore, the student should be allowed partial or full credit for explanations which bear up under scrutiny. On the surface such an approach sounds at best unworkable, at worst indefensible. The approach may seem unworkable because of the long hours its administration would seem to entail. The approach may seem indefensible because of the door it opens to the monster of subjectivity. I hesitate even at this moment to go forward with the attack But the end seems more than worth the ridicule that may stand in the way Note that I am not alone. Individuals who denounce the present standardized tests almost to a person make the same criticisms of the multiple-choice format that I have made. By implication, they must either go on to argue against all testing whatsoever, or suggest some alternative approaches Unfortunately, no generally attractive alternatives have rolled off the pens of today's critics. So I am left to breach the gap between the hoped for and the possible.

I have chosen to argue for the multiple-choice format as a way of testing functional literacy skills. However, I have added an important proviso. I have suggested that the student be allowed to defend his answers as he proceeds through the tests. It is now time for me to offer a brief defense of my own choices. At least four arguments suggest themselves. First, the experience is an instructive one for teachers and other individuals involved in the administration of such an exam. Presumably the teacher is accountable for behaviors on the test. It is to his or her advantage to become as familiar with the important areas of functional literacy as possible. On the one hand, the teacher involved in the type of testing I am suggesting is confronted with deciding what it is that constitutes the general nature of correct and incorrect answers. On the other hand, the teacher becomes more aware of the students' strengths and weaknesses through listening to the students' responses. Second, such a way of testing still retains a fair share of objectivity. The item stem, the item itself, and the alternatives are identical for each and every student. Third, the student as well as the teacher stands to gain from such a procedure. The construction and defense of an argument in the space of a few minutes is a skill to be valued in itself. But fourth and perhaps most important, the increase in the construct validity of such a test over traditional tests seems inevitable. So, the procedure I have sketched seems to be of benefit to everyone. In our rush to avoid subjectivity we may have lost sight of the importance of construct validity. With more objectivity came a decrease in construct validity. Perhaps it is high time that a more favorable balance was struck.

Concurrent Validity

The proliferation of techniques used to place individuals into the category of master and nonmaster is bewildering at best, and counterproductive at worst. All the techniques rest on the assumption that mastery and non-mastery are meaningful concepts in the domain being tested. It is intuitively plausible that such areas as mathematics and the sciences may satisfy this assumption. However, the generally arbitrary nature of cut-off scores has proved so troubling to some people that they seriously question the merits of determining and using cut-off scores at all (Hambleton, 1978, Glass, 1978(2)). Nevertheless, we will assume for the moment that one can legitimately divide the relevant portions of the world into masters and nonmasters. The approach to a cut-off score determination has until very recently been largely based on some form of agreement between experts in the field (see Meskauskas, 1976, for a good review of both this approach and the following approach). Recently this has been augmented by helpful statistical techniques. I will speak briefly of both approaches.

One of the first attempts to arrive at a cut-off score was undertaken in the late 1940's for a University of Chicago departmental physics course (Nedelsky, 1954). The department, which taught physics courses by means of a common

subject outline, generated a joint departmental comprehensive examination consisting of over 200 five-choice questions. Each of the approximately six instructors who were teaching sections at a given point in time were asked to look at the test prior to the candidate's taking it. The instructors had to decide for each question which of the distractors the lowest passing student (in this case a D Student) should be able to identify as incorrect. The minimum passing level, or MPL, for each item is the reciprocal of the number of remaining alternatives. For example, if in a five-choice question only one of the distractors is marked as one that the lowest passing student should be able to eliminate, the minimum passing level for the item is ¼ since there are four remaining alternatives. Each question was rated by all the instructors in this manner. For a five-choice item the possible values are .20, .25, .33, .50, and 1.0. The MPL for the examination consisted of a summation of these individual item MPL values. The method becomes a bit more complicated in practice, but the above reflects the general idea well enough. Figure 9 shows how one might arrive at a cut-off score for an exam with five choices or alternatives, five items and three judges.

Several criticisms of such a method quickly come to mind (Meskauskas, 1976). One of these criticisms is a starting point for statistical procedures to cut-off scores. Note that errors of measurement did not enter anywhere into the discussion of the above method. However, in 1971 Emrick noted that measurement errors would cause a number of non-masters to be included in the master category. These were called alpha errors or false positives. Similarly, a number of masters would be included in the non-master category. These were called beta errors or false negatives. In many cases we might like to know the relative abundance of false positive and false negative errors. Furthermore, we might well want to change cut-off scores so that they gave more weight to the false negative errors than they gave to the false positive errors. That is, we might consider it very important to classify all masters as such. Emrick's particular solution to this problem has since been disputed (Wilcox and Harris, 1977). However, the importance of being able to distinguish between alpha and beta errors is still with us and has worked its way into a number of other models for determining cut-off scores. (See, for example, Fhaner, 1974).

Predictive Validity

Finally, we come to a consideration of the efforts taken to increase the predictive validity of measures of functional literacy, to increase the extent to which success on the measure predicts success in adult life. There are in general two ways of going about the task. Both ways can be illustrated by the work done at the University of Texas in what has come to be called the Adult Performance Level Project (Adult Performance Level Project Staff, 1973). First, the predictive validity of a measure may be increased in a relatively simple and straightforward way. If experts can agree on what the minimally literate adult must be able to read, then these experts' opinions can be put to

good use. Items can be placed on the measure which reflect the experts' opinion of what must be tested. A reasonably conscientious and careful project can go a long way toward clarifying the content of what must be tested as well as increasing the potential predictive validity of the measure. Before confusion sets in, let me distinguish between the concern with content validity in an earlier section and the concern with predictive validity in this section. In the section on content validity I assumed that the general domain of importance had already been specified. The job was to fill in the domain with the appropriate content. Here I am not assuming that the general content area has been specified. Thus we are backing up a step.

The manner in which the Adult Performance Level people set out the areas of importance is worth noting (Figure 10). They divided the world into general knowledge areas and basic skills. There were six general knowledge areas. occupational knowledge, consumer economics, health, community resources, government and law, and transportation. And there were six basic skills, reading, writing, listening, computation, problem solving, and interpersonal relations. My point in bringing up this example is not to suggest that there is anything particularly good about their division of knowledge areas and basic skills. My point is more general. A matrix such as the one which appears in Figure 10 allows one to be complete, to forge ahead with some map of the universe. I think such a map is a welcome adjunct to our intuitive notions of what are and what are not minimal competencies.

There is a second way one might go about increasing the predictive validity of measures of functional literacy. One might worry less about what the

Item	Judge 1		Judge 2		Judge 3		Item MPL
	A[1]	B[2]	A	B	A	B	
1	2	1/3	3	1/2	2	1/3	.39
2	4	1	3	1/2	4	1	.83
3	1	1/4	1	1/4	0	1/5	.23
4	2	1/3	1	1/4	2	1/2	.36
5	3	1/2	3	1/2	2	1/2	.25

Test Minimum Passing Level: (.39+.83+.23+.36+.25) = 2.06

[1]A. # of choices minimally competent student should be able to discard

[2]B: reciprocal of # of remaining items, i.e., expected value for an item with equiprobable choice of remaining items

FIGURE 9

minimally competent person ought to read and more about what the minimally competent person should have achieved in his job and other life activities. The only thing that justifies the previous procedure is the notion that the materials we select are indeed needed to be competent. Instead of this more or less subjective approach, one might identify various indices of competence. That is, a person who is placed high on an index of competence should score well on an item designed to measure competence skills. Again, this is just what the Adult Performance Level people did. They identified three indices of competence, occupational prestige, education and weekly income. Four levels to each index were defined. Scores on an item were then correlated with level of an index (Figures 11 through 13). Items on which scores correlated well with the level of an index were kept in, other items were thrown out.

Lest you think everything is turning up roses the following comment by a well-respected educator is in order, (Glass, 1978 (2)):

> To my knowledge, every attempt to derive a criterion score is either blatantly arbitrary or derives from a set of arbitrary premises. But arbitrariness is no bogeyman, and one ought not to shrink from a necessary task because it involves arbitrary decisions. However, arbitrary decisions often entail substantial risks of disruption and dislocation. Less arbitrariness is safe.
>
> Teachers and their consultants attempting to define "competence" and writing test items intended to reflect minimal levels of acquisition are engaged in a bootless and potentially embarrassing endeavor. They are likely to construct a competency-based test for graduation that is quite inappropriately difficult. Then they will be forced to back off and will be accused publicly of either not knowing what students ought to know or else not teaching students what they ought to learn. They are in fact guilty on neither account. No one knows how well a person must read to succeed in life, or what percent of the graduating class ought to be able to calculate compound interest payments.

I must confess that I agree with the spirit of Dr. Glass's remarks. It does indeed seem to me rather naive to assume that we can actually find a set of items which more or less guarantees success or failure in life. Perhaps the whole notion of minimal competencies is as silly and as useless as the vote taken during one's senior year in high school on the student most likely to succeed. But these remarks fall generally outside the substance of this talk. For our purposes we need only note that one can seemingly take measures that improve the predictive validity of our instruments.

Conclusion

It should be clear by now that many standardized tests are simply very poor measures of functional literacy. One can all too easily find fault with the

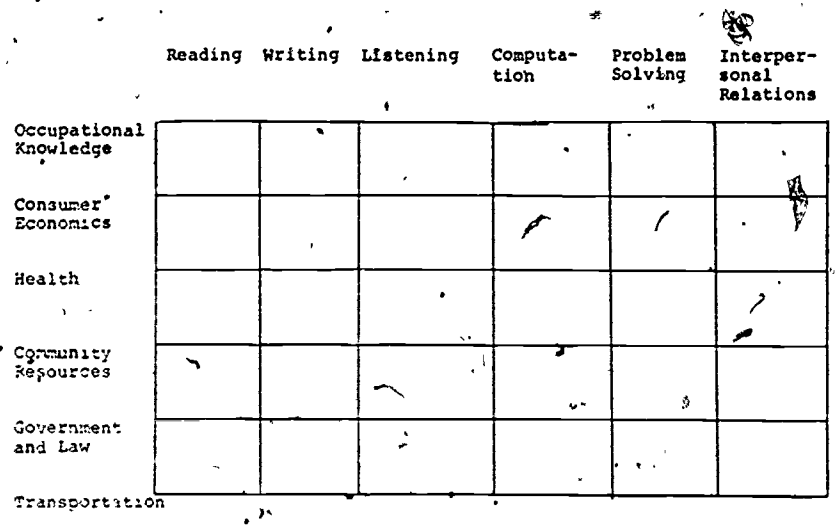


FIGURE 10

validity of most standardized measures. And valid is just what we want our measures to be.

Implicit in my criticism of standardized tests have been the following two equally important, if not more important, criticisms. First, it is simply bad economics to go on testing as we do. Too many children may be misclassified. Teachers will learn little if anything from the testing situation. Diagnosis and remediation may well be unrelated to the real problems of students. But bad economics as a criticism pales before the inherent unfairness of standardized tests. Students are sentenced to a test score without a trial. Students are not allowed to defend their answers. If an answer is given which test developers did not consider, so much the worse for the student. There is no reason for tests to be the arbitrary, often capricious dictators of students' lives that they are. Something can be done. Something should be done, for we are no longer innocent.

REFERENCES

Adult Performance Level Project Staff. The Adult Performance Level Study. Austin, Texas. Division of Extension (University of Texas), January 1973.

Anastasi, Anne. Psychological Testing. New York. MacMillan, 1965.

Bormuth, John R. Reading literacy. its definition and assessment Reading Research Quarterly, V IX N1, 1973-1974, pp. 6-7.

Brickell, Henry M. Seven key notes on minimum competency testing. Phi Delta Kappan, V 59, N 9, May 1978, pp. 589-592.

Cawelti, Gordon National competency testing. A bogus solution. Phi Delta Kappan, V 59, N 9, May 1978; pp. 619-620.

Cronbach, L J Test Validation. In R L Thorndike (Ed). Educational Measurement (2nd edition). Washington, D.C.: American Council on Education, 1971.

Ebel, R L Content standard test scores Educational and Psychological Measurement, V22, 1962, pp. 11-17.

Ebel, Robert L The case for minimum competency testing Phi Delta Kappan, V59, N8, April 1978, pp 546-548.

Emrick, J A An evaluation model for mastery-testing Journal of Education Measurement, V8, 1971, pp 321-326.

Engel, Brenda One way it can be. Today's Education, V66, N2, March April 1977, pp 50 52.

Explanation of multiple-choice testing Princeton, N J Educational Testing Service, 1961

Item 9

If you wanted to apply for the job shown below, which of the following application methods would you use?

Security Officers--Start $2 per hour, uniforms furnished. Apply 801 W. 24th St., after 6 p.m. Holmes Lobby Desk.

a. telephone call
b. written application (resume)
Xc. in person application
d. I don't know

Eighty-two percent of the sample answered this item correctly, while 13 percent answered incorrectly. Percent correct responses according to criterion variables are given in Figure 9.

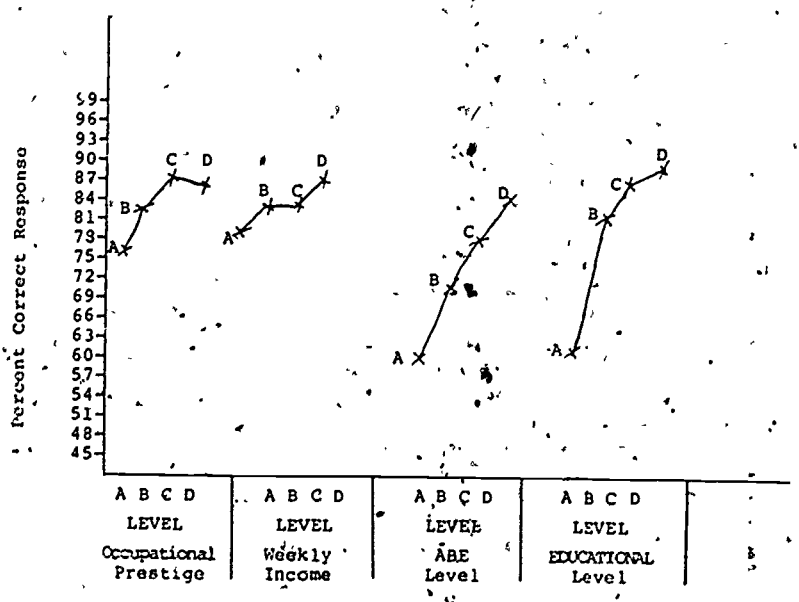


Figure 9. Occupational Knowledge referenced item on work: application procedure

FIGURE 11

Farr, Roger. Reading. what can be measured (IRA Research Fund Monograph). Newark, Delaware: International Reading Association, 1969.

Fremer, John. In response to Gene Glass. Phi Delta Kappan, V 59, N 9, May 1978, pp. 605 606.

Glass, Gene V. Minimum competence and incompetence in Florida. Phi Delta Kappan, V 59, N 9, May 1978, pp. 602-604.

Glass, Gene V Matthew Arnold and minimal competence. Educational Forum, January 1978, pp. 139-144.

Hambleton, Ronald K., H. Swaminathan, J. Algina, D.B Coulson. Criterion-referenced testing and measurements, a review of technical issues and development Review of Educational Research, V 48, N 1, Winter 1978, pp. 1-48.

Hively, W., H.L. Patterson, S.A. Page. A 'universe-defined' system of arithmetic achievement tests Journal of Educational Measurement, V 5, 1968, pp 275 290

Item 4

Mr. Packard wants to buy a car. The salesman says that he can pay for it over a year and that, plus interest, the price will be $255.66. Interest is:

a. the salesman's salary
b. the actual value of the car
Xc. the cost charged for handling the deal on the time basis
d. a state tax
e. I don't know

The percent correct response attained by the APL sample on this item was 70 percent. This item, like the preceding ones, also, dealt with a commerical term (interest). However, the item differentiated among the levels on all four group variables. The more successful persons were more likely to know the meaning of interest. Percent correct responses for the criterion variables are presented in Figure 4.

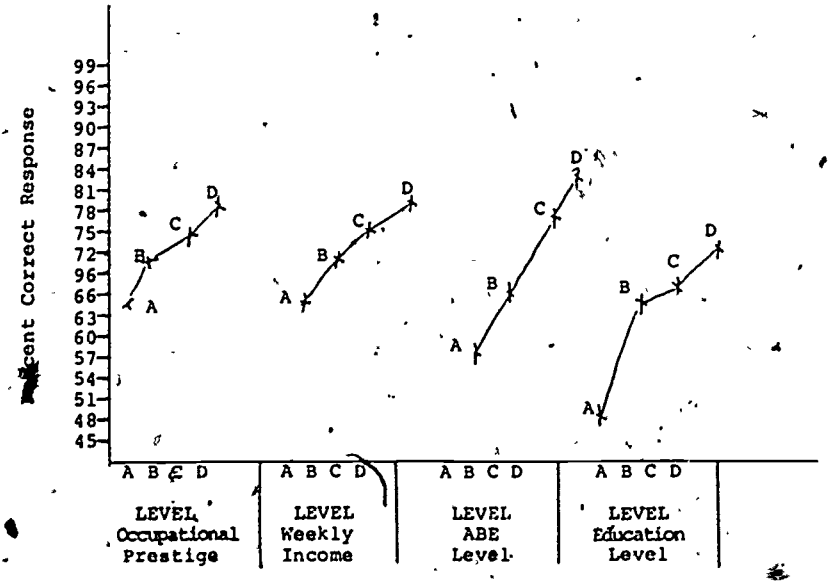


Figure 4. Consumer economics referenced item on commercial term:

FIGURE 12

Hoffman, Banesh The tyranny of testing. New York. Crowell-Collier, 1962.

Jennings, W and J Nathan Startling, disturbing research on school program effectiveness. Phi Delta Kappan. March 1976. pp. 568 572.

McKenna Bernard What's wrong with standardized testing Today's Education. March April 1977, p. 36

Meskauskas, John Evaluation models for criterion referenced testing views regarding mastery and standard setting Review of Educational Research. V 46. N 1. Winter 1976. pp. 133 158.

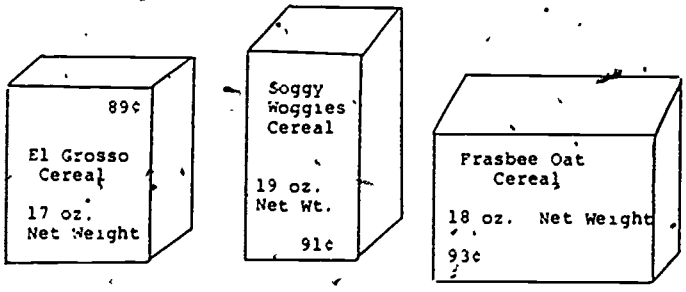
Murphy, Richard T Adult Functional Reading Study (PR 75 2) Princeton. N.J. Educational Testing Service. 1975

Nathan. Joe and Wayne Jennings Educational bait-and-switch Phi Delta Kappan. V 59, N 9. May 1978. pp 621 625

Items 31-32

This two-item exercise was developed to test the adults' ability to calulate weight and price per unit in order to arrive at the most economical buy on food purchases.

Directions: Below you will see three boxes of cereal. On each box is printed the price and the weight. Look at the prices and the weights and then answer the two questions below, please.



Item 31

Which of the three boxes of cereal is the best buy?

Answer: Soggy Woggies Cereal

Only 52 percent of the sample answered this item correctly. Chi square values of pattern of response reached significance on three of the criterion variables. Weekly Income was the exception. Adults in Level C of Occupational Prestige ratings did slightly better than adults in the higher Level D rating. The overall trend of response was in an ascending pattern, indicating that adults in the higher levels were more successful in figuring which cereal was the best buy. Figure 12 gives percent correct responses by levels across the criterion variables. (See Figure 12).

Item 32

Which of the three boxes of cereal contains the most cereal by weight?

Answer: Soggy Woggies Cereal

FIGURE 13

National Education Association Task Force on Testing. A summary of alternatives. Today's Education. V 66, N 2, March April 1977, pp. 54 55

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

Perrone, Vito. On standardized testing and evaluation. In Paul L. Houts (ed). The myth of measurability. New York: Hart Publishing Company, 1977

Pipho, Chris. Minimum competency testing in 1978: a look at state standards. Phi Delta Kappan, V 59, N 9, May 1978, pp 585 588

Popham, W James. An approaching peril: cloud-referenced tests. Phi Delta Kappan, V 55, N 9, May 1974, pp 611 614

Regents of the University of California, Davis V Allan Bakke. Pacific Reporter 2nd, v 553, p 1151

Vineberg, R , T G Stitch, E N Taylor, J S Caylor. Effects of aptitude, job experience and literacy on job performance: summary of HumRRO work units UTILITY and REALISTIC (FR 71 1) Alexandria, Va Human Resources Research Organization, February 1971

Taylor, Edwin F. The looking-glass world of testing. Today's Education, March April 1977, pp 39 44

Wilcox, R and C W Harris. On Emrick's "An evaluation model for mastery testing." Journal of Educational Measurement, V 14, N 3, Fall 1977, pp 215 218

U S GOVERNMENT PRINTING OFFICE 1980 O 302 477