DOCUMENT RESUME

ED 227 150 TM 830 161

AUTHOR Bliss, Leonard B.
TITLE Validation of the Use of the Stanford Achievement Test with U.S.V.I. Students. Virgin Islands of the United States Public School Basic Skills Assessment Survey, Technical Report No. 1.
INSTITUTION College of the Virgin Islands, St. Thomas. Caribbean Research Inst.
PUB DATE Jan 82
NOTE 54p.; Paper presented at the Annual Meeting of the American Educational Research Association (66th, New York, NY, March 19-23, 1982).
AVAILABLE FROM Caribbean Research Institute, College of the Virgin Islands, St. Thomas, USVI 00802 ($2.00).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Achievement Tests; *Basic Skills; Data Analysis; Educational Assessment; Elementary Secondary Education; Language Skills; Mathematics Achievement; Reading Achievement; Sampling; *Standardized Tests; *State Programs; *Test Reliability; *Test Validity
IDENTIFIERS *Stanford Achievement Tests; Virgin Islands

ABSTRACT

A sample of slightly over 1500 students was drawn from even-numbered grades in public schools of the U.S. Virgin Islands, and was given the 1973 edition of the Stanford Achievement Test (in grades 2,4,6, & 8) and the Test of Academic Skills (grades 10 and 12) to assess student academic achievement in the basic skill areas of mathematics, reading, and English language. This report describes phase I of the data analysis, which involved the determination of levels of content validity and reliability of the scores obtained from these Virgin Islands students on these tests which were originally standardized on continental United States populations. The results indicate that the tests are content valid for use in Virgin Islands public schools at these grade levels and that the scores obtained are at least as reliable as those obtained using continental U.S. students during the test standardization procedures. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED227150

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy

Virgin Islands of the United States Public School Basic Skills Achievement Survey

Technical Report #1: Validation of the Use of the Stanford Achievement Test With U.S.V.I. Students

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

L.B. Bliss

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

PRESENTED AT THE ANNUAL MEETING
OF THE AMERICAN EDUCATIONAL
RESEARCH ASSOCIATION
NEW YORK CITY, MARCH 19-23, 1982

TM 830 161

Caribbean Research Institute, College of the Virgin Islands
Leonard B. Bliss, Ph.D. - Principal Investigator

January 1982

PREFACE

With the appearance of Virgin Islands of the United States Public School Basic Skills Achievement Survey, Technical Report #1: Validation of the Use of the Stanford Achievement Test With U.S.V.I. Students the Institute has embarked on the publication of a Working Paper Series. These papers are intended to present the author's (and the Institute's) point of view on various subjects as a matter for discussion and comment by those who agree as well as disagree with expressed positions. In this way the Institute hopes that the final versions will be improved in style as well as rigour.

The present paper is the first phase of a study of basic skills in the schools of the United States Virgin Islands requested by the Board of Trustees of the College. The work has taken considerably longer than anticipated due to fundamental alterations in the design so as to provide greater depth than originally planned. Unfortunately, shortage of staff did not allow the progress hoped for to be made.

The data for the whole project have been collected, however, and work is proceeding on their interpretation and the compiling of the three reports which will follow.

Norwell Harrigan
Director

Abstract

A sample of slightly over 1500 was drawn from even numbered grades in public schools of the U.S. Virgin Islands and were given the 1973 edition of the Stanford Achievement Test (in grades 2,4,6, & 8) and the Test of Academic Skills (grades 10 and 12) in an attempt to assess student academic achievement in the basic skill areas of mathematics, reading, and English language. This report describes Phase I of the data analysis which involved the determination of levels of content validity and reliability of the scores obtained from these Virgin Islands students on these tests which were originally standardized on continental United States populations.

The results indicate that the tests are content valid for use in Virgin Islands public schools at all of these grade levels and that the scores obtained are at least as reliable as those obtained using continental U.S. students during the test standardization procedures.

ERIC

It is almost becoming a matter of faith that achievement in basic skills (i.e. English language and mathematics) in public schools under the American flag has deteriorated over the last twenty years. Proponents of this idea point to evidence as formal as decreases in typical scores on the Scholastic Aptitude Test and standardized tests of academic achievement and as informal as the quality of writing and arithmetic skills they perceive in the young people around them.

The reactions of people to this perceived phenomenon are also varied. On the government level they include the requirement that all students score a minimum grade on a test of basic skills in order to receive a high school diploma; that teachers pass a similar test to obtain teacher certification; and that schools require students to take additional course work in basic skills areas. In addition, federal, state, and local governments have initiated programs to provide support in the forms of grants and technical assistance to schools at all levels to do research and set up programs designed to improve student achievement in basic skills.

At a different level, parents, concerned that the public schools are not doing an adequate job in preparing their children in basic skills areas, are choosing, in increasing numbers, to remove their children from public schools and place them in religious and secular private schools. While there are other reasons for the proliferation of private schools besides the purely academic, the desire for high quality academic preparation is one compelling cause of this phenomenon.

The public schools, themselves, have reacted strongly to this crisis in public confidence. These reactions include an increase in required courses in language and mathematics areas with a corresponding decrease in electives in areas considered less "basic." Projects to revise curricula in basic skills areas proliferate and are receiving more support than they have since the reevaluation of American education engendered by the shock of Sputnik in the late 1950's.

Improving basic skills achievement was a concern of the Department of Education of the government of the Virgin Islands of the United States when it approached the College of the Virgin Islands to provide aid in improving such instruction. In an effort to provide this service, the Caribbean Research Institue, the college's research arm, worked with a task force composed of representatives from the Department of Education and CRI to determine a course of action.

It became clear after the first few task force meetings that development of any strategy designed to improve basic skills achievement needed to start off with a fairly detailed description of current achievement levels of students in territorial public schools. This information was not available. Public school students were administered a standardized achievement test only at the end of sixth grade (The Iowa Test of Basic Skills). In other elementary grades most students were tested annually or semiannually at their schools, but the test given and the times during the academic year that were administered varied greatly and apparently at the whim of building administrators. The results of these tests stayed at the schools and were not collected at any central point. On the secondary level there was no program of standardized achievement testing.

An additional factor which limited the use of previously collected achievement level data was that all scores were reported in a norm referenced manner. That is, scores did not indicate which basic skills examinees had or lacked, but rather how examinee's scores compared to those obtained by a group of students to whom the tests were previously administered in the continental United States. The Iowa Test of Basic Skills administered to sixth graders did make comparisons with other V.I. sixth grade students (i.e. they reported using local norms), but even these were of no use in determining whether or not individual students had attained specific basic skills.

It was decided to test a representative sample of U.S. Virgin Islands public school students using a standardized basic skills battery. Choosing the test, the following criteria were used:

1) The test must be technically sound in terms of reliability and item discrimination, at least for the group it had been field tested on.
2) The test must be content valid for U.S. Virgin Islands public school students. That is, there

needed to be a high degree of matching between the content and behaviors sampled by the test and those actually in the curriculum taught at various levels in the U.S.V.I. public schools.

3 The test must include a detailed statement of the objectives tested while providing an item by objective keying procedure.

4) Scores which indicate students' performances relative to each objective must be available. That is, criterion-referenced scoring must be provided.

The 1973 version of the Stanford Achievement Test (Basic Battery) was chosen as the test which appeared to meet the criteria listed above. It was administered to slightly over 1500 students in the Fall of 1980 in both the St. Thomas/St. John and the St. Croix school districts. This is the first of a series of research reports designed to make available the results of this rather complex study. A simple, brief example of the quantity of data obtained may serve to highlight the scope of this study. The Intermediate Level II of the Stanford Achievement Test (administered to sixth graders in this study) contained 351 items. It was administered to 225 students in the U.S. Virgin Islands sample yielding 78,975 individual pieces of data. The sixth grade sample, due to a technical difficulty (the principal in one school forgot to assign the teacher of the selected class the task of giving the test and the teacher in another school administered only four of the seven subtests), contained the smallest number of examinees of any grade level. Additional reports will be issued regularly as soon as results become available.

ERIC

Validity and Reliability of Test Scores

This first report deals with the establishment of the validity and reliability of the test scores. Validity refers to the extent to which the test measures those characteristics which it is intended to measure. Ebel (1961) has referred to validity as "one of the major dieties in the pantheon of the psychometrician" (p.640). Three types are now commonly used in educational and psychological measurement (see French and Michael, 1966). These are content, criterion-related, and construct validity.

Gronlund (1976) indicates that, "Criterion-related validity may be defined as the extent to which test performance is related to some other valued measure of performance" (p. 83). This may be performance on a task in the future (i.e. predictive validity) or on some present objectives not directly measured by the test (i.e. concurrent validity). Since the purpose of administering an achievement test is to get a <u>direct</u> measure of <u>present</u> student mastery of certain academic objectives (i.e. there is no attempt to predict future performance or to infer performance levels on objectives not directly measured by the test), criterion-related validity is not an issue in determining the appropriateness of the Stanford Achievement Test in measuring academic achievement in this study.

The term "construct validity" was first introduced into the area of psychometrics by Chronbach and Meehl (1955) who defined a construct as a postulated (that is, assumed or hypothetical) attribute of people that underlies and determines their overt behavior. If the behavior can be directly observed, or if the trait can be operationally defined, it is not a construct in this sense. Ebel (1979) notes

> Most of what we teach in educational institutions are knowledges, skills, and abilities. These can all be defined operationally. They are not hypothetical constructs. Ability to type, to spell, to weld, ability to solve problems with algebra, calculus, or computers; these are not the kind of latent traits Cronbach and Meehl had in mind. We would speak more sensibly, I think, if we did not call them constructs. (p. 307)

Construct validity is concerned with whether or not a test is accurately measuring the construct it purports to measure Since this study is operationally defining basic skills achievement as the performance of students on the Stanford Achievement Test, it is clear that no construct is being measured. Hence, construct validity will not be a concern in this report.

The content of any curriculum can be thought of as being composed of subject matter content and behavioral changes sought in students. For a test to be content valid it must provide results that are representative of the topics and behaviors we wish to measure. More formally, ". . . content validity may be defined as the extent to which a test measures a representative sample of the subject matter and the behavioral changes under consideration" (Gronlund, 1976, pp. 81-82). Effective strategies for determining content validity involve determining the objectives sampled by the test and examining the curriculum to ascertain the degree of match between them. Achievement tests are primarily concerned with measuring the acquisition of certain skills and knowledges (objectives) by students at the time that the test is given. Thus, it is content validity that should be of prime concern in this study. Specifically, do the objectives tested by the Stanford Achievement Test correspond to those taught toward in the schools of the U.S. Virgin Islands?

Reliability deals with the consistency of the <u>scores</u> of a test over time and over different examinees. It is purely a statistical phenomenon and cannot be determined logically as can content validity. Furthermore, it is a function of the <u>scores</u> of the test rather than of the test, itself. This means that a test may give highly reliable scores for one group of examinees, but result in lower reliability with another group. In essence, what we are concerned with is whether or not the test scores represent measures of the same traits each time the test is given.

It is important, then, that whatever measure of basic skills achievement is used, that the measure be content valid for the curriculum used in Virgin Islands public schools and produce reliable scores when administered to Virgin Islands public school students.

As any good commercially available standardized test, the 1973 edition of the Stanford Achievement Test was standardized on a large sample of students. The SAT <u>Technical Data Report</u>

ERIC

(1975) indicates that a sample of over 275,000 pupils from 109 school systems in 43 states in the United States made up the standardization samples used. Table 1 provides descriptions of these samples and how they compare with a description of the population of the continental United States. Content validity was established by curricular analysis using information from a large number of sources.

> Basic to the construction of a series of achievement tests is the identification of what is being taught in the schools across the nation. The most important sources for curricular analysis were (a) textbook series in various subject areas (including the preparation of detailed analysis of the content of the books most widely used in each field); (b) a wide variety of courses of study from individual school systems; (c) statements of objectives from various state and national committees, and the opinions of experts in various fields; and (d) the research literature pertaining to children's concepts, experience, and vocabulary. (Technical Data Report, p.12)

The reliability of the scores of the standardization sample was determined by using the Kuder-Richardson Formula 20 and by calculating the standard error of measurement of the scores. Two measures of reliability were used since it is known that high homogeneity in tested groups will lower the reliability estimates obtained using the KR-20, but that his effect is dealt with in determining standard errors. In addition, the standard error of measurement is more meaningful in interpreting scores of individual students. With very few exceptions, the reliabilities obtained from the standardization samples ranged from .84 to .95 using the KR-20 formula.

While the 1973 version of the Standford Achievement Test appears to be educationally sound based on the standardization

Table 1

Summary of Characteristics of Standardization Samples[1]

Characteristics	Stanford Population	Stanford Range	National U.S. Population 1970 Data
Percent of pupils by community size			
0-49, 999	70.0		64.1
50,000-249, 999	14.2		15.2
250,000 or more	15.9		20.7
Percent of pupils by Geographic Region			
Southeast	23.8		22.2
North Central	21.6		27.8
Northeast	26.5		24.2
West	28.2		25.8
Median Family Income	$9,096	$ 4,878 to $13,593	$9,590
Median Years of Schooling (Adults 25 yrs. & older)	12.1	8.4 to 12.6	12.1
Average Class Size (Student-Teacher Ratio)	26.4	18 to 36	24.3
Average Starting Salary of Teachers	$7,116	$ 4,500 to 11,500	$7,064
Average Salary of Teachers	$9,360	$ 4,500 to 11,500	$9,265
Median Years Teaching Experience	10.8	5 to 24	10
Percent of Grade 1 pupils who attended kindergarten	84.6	0 to 100	71.8
Percent of Schools Using Some Team Teaching	67.1		

Table 1 continued

Characteristics	Stanford Population	Stanford Range	National U.S. Population 1970 Data
Percent of Schools Using Some Teacher Aids	97.5		
Percent of Pupils Not Promoted to Next Highest Grade			
Grade 1	3.9	0.0 to 25	
Grade 2	1.8	0.0 to 15	
Grade 3	1.5	0.0 to 10	
Grade 4	0.9	0.0 to 10	
Grade 5	0.8	0.0 to 5	
Grade 6	1.1	0.0 to 5	
Grade 7	1.2	0.0 to 11	
Grade 8	1.3	0.0 to 9	
Grade 9	2.4	0.0 to 9	
Percent of Pupils n Non-public Schools	9		12
Percent of Major Ethnic Minorities			
Blacks	11.6	0 to 60	11.1
Hispanics	4.6	0 to 60	4.6
Other	Less than 1		Less than 1

[1]From Stanford Achievement Test: Technical Data Report, p. 21.

groups data, the groups contained only continental U.S. students. Likewise, the test makers most probably did not take Virgin Islands public school curriculum into account when designing items. Therefore, before the scores of any tests of basic skills can be used to draw conclusions about V.I. students, the content validity and reliability of these test scores for Virgin Islands students must be established. Hence, this report.

ERIC

Method

Sampling

The June 1, 1979 enrollment in the public schools in the Virgin Islands of the United States was 25,426 according to the statistics issued by the V.I. Department of Education. It was clear that testing this number of students was economically unfeasible. The preferred alternative would have been to generate a random sample of students in grades K-12 to be tested, but it was equally clear that this would have produced an intolerable disruption of classroom activities. Therefore, in an attempt to obtain a representative sample of students, cluster sampling was used with the clusters being defined as classes. The number of classes to be selected for the sample from each grade in each of the St. Thomas/St. John and St. Croix districts was determined by calculating the proportion of the total K-12 student population in each grade in each district and assuming a class size of thirty.

Selecting whole classes presented an additional difficulty. The small number of classes selected in each grade might have made obtaining a representative sample of students more difficult. This is due to the fact that while classes in a given elementary school may be heterogeneous, the schools themselves are not. This is because elementary schools in the U.S. Virgin Islands are essentially neighborhood schools. Virgin Islands neighborhoods tend to be homogeneous in terms of socioeconomic status of residents. To overcome this problem, it was decided to increase the number of classes tested in a given grade (thereby increasing the number of schools within the territory from which these classes came) without increasing the total number of students tested by testing at alternate grades. This seemed acceptable since many of the objectives tested by the Stanford Achievement Test carry across adjacent levels of the test and there was no reason to suspect that the patterns of academic achievement of students in odd numbered grades were different from those in even numbered grades.

It was originally proposed that students in odd numbered grades be tested during the Spring of 1980, but difficulties in obtaining testing materials resulted in testing being postponed until the Fall of 1980. In order to deal with the cohort of students originally selected, even numbered grades were actually tested.

The classes to be tested were chosen by chance. Specifically, for each grade in each district a listing of classes was made and each class was assigned a number. A table of random numbers was consulted. Numbers were drawn from the table until there were the same number of random numbers chosen as there were classes needed for the sample. In the case of duplicate numbers being drawn, the duplicate was ignored and another number chosen. If the number chosen was outside the range of the number of classes on the list, it was ignored and another number was chosen. When sufficient numbers had been drawn, the listed classes which corresponded to these numbers were included in the sample. This procedure was repeated for each grade in each district.

The sole exception to this procedure was in the eighth grade portion of the sample. On St. Thomas, homeroom classes are somewhat homogeneous in that students repeating eighth grade and those in the eighth grade for the first time are placed in separate homeroom classes. Since the levels of academic achievement for repeaters and nonrepeaters are very likely different, the proportion of repeaters and nonrepeaters was determined to come out with a number of classes needed in the sample from each group and the groups of classes of repeaters and nonrepeaters were sampled separately in the manner described in the preceding paragraph.

Elena Christian Junior High School on St. Croix is on split session. The principal of that school felt that there were definite differences in achievement levels between the students in the morning and afternoon sessions. Because of this, classes in the morning and afternoon sessions were sampled separately using the same procedure employed on St. Thomas for the repeating and nonrepeating homeroom classes.

If simple random sampling has been used in selecting students to be tested, a sample size of approximately 2000 would have been the maximum size required to obtain an accuracy of about ±2% at a .95 level of confidence when estimating the proportion of V.I. students reaching certain objectives from the sample proportions if a typical proportion answering each item correctly were .50 (see Asher, 1979, p.166). In actuality, due to student absences, failure of school personnel to carry out requested tasks, and other difficulties, the sample size obtained was only 1535. However, examination of the difficulty indexes of the Stanford Achievement Test items on all levels revealed difficulty indexes considerably different from .50 on most items. This would tend to shorten the size of the confidence interval. Finally, the financial and organizational constraints cited previously forced the investigators to use cluster sampling techniques rather than random sampling. Since the intraclass correlations (i.e. the effects of clustering on the standard deviations of the achievement test scores) were not known, this factor also contributes toward making the above mentioned accuracy estimate a rather crude one. It can, however, serve as a rough guideline.

Table 2 presents the relevant sample size data. The sixth and second grade samples from St. Croix are smaller than had been hoped for the following reasons. As indicated previously, the teacher of one of the sixth grade classes only administered four of the seven subtests. In a second grade class, the teacher was ill during the days set aside for testing and the test was not administered. By the time this became apparent to the investigators, it was too late to go back to St. Croix to retest.

Aside from the difficulty in estimating precision of the proportions of students obtaining correct scores on various items, the sampling procedure used presents another difficulty. Because of the previously stated practical considerations, it was necessary to employ cluster sampling (sampling whole classes) rather than simple random sampling of students to be tested.

Table 2

U.S. Virgin Islands Sample Sizes

Grade	Test Level	Total System	St.Thomas/St.John District	St. Croix District
12	TASK II	129	74	55
10	TASK I	254	167	87
8	Advanced	345	173	172
6	Intermediate II	227	146	81
4	Primary III	346	186	160
2	Primary I	234	143	91
	TOTAL	1535	889	646

The principal drawback to cluster sampling is the likelihood of increased sampling error. In general, as the size of the sample increases, the size of the standard error decreases. This applies, however when each sample element (in this case, each student) is selected independently of every other element. In cluster sampling the elements are, by definition, selected in a group rather than independently. The effect of clustered selection on the standard error will depend on the similarity between the elements in the cluster and those in the population. In many cases, sample elements selected in clusters will not show the same variation as an equivalent number selected independently. Students who attend the same school and are in the same class may be more like one another in a characteristic such as academic achievement than students in the public school population as a whole.

The relationship between clustering and sampling error may be summarized as follows. If all the elements (students) in a cluster (class) were identical with regard to achievement and totally different from the elements in other clusters, the sampling error would be extremely high. Clustering, in this case, would tend to make the clustered sample equivalent in size to a simple random sample with as many subjects as there are clusters, rather than elements. Hence, a sample made up of 60 clusters might be equivalent to a simple random sample of 60 individuals. This is obviously an extreme case that is never seen in practice. At the opposite extreme would be a series of clusters showing the same variation within each cluster as simple random samples of the same size. In this case, each cluster would represent the entire population, another condition rarely met in practice. Most sampling situations fall in between these extremes, tending toward one or the other according to the characteristic being studied. In general, according to Warwick and Lininger (1975), experience has shown that well-designed cluster samples will produce standard errors that are about one and one-half times as large as the standard errors from simple random samples of the same size.

This situation should not have any effect on the descriptive statistics reported in this document, but it will enter into the interpretation of the results of hypothesis testing using parametric techniques since these latter techniques rely on estimates of the standard error. These will be discussed as the results of these tests are dealt with. In general, however, the resulting under-estimates of the standard error of the means will result in test statistics that are higher than they would have been if clustered standard error estimates had been used.

Notwithstanding the difficulties involved in sampling, the researchers are confident that the resulting samples are as representative of the entire V.I. public school population as is really possible given the nature of working with human subjects and the organizational considerations of schools combined with the resources available for the study. The difficulties encountered are not untypical of those commonly found when doing field work in both public and private schools.

Testing Procedure

Testing was done at the grade level recommended by the test publisher. Table 3 indicates the subtests of the battery given to each grade. This was primarily done to insure the content validity of the examinations. Tests were administered by classroom teachers or guidance counselors, at the discretion of building administrators. Each person who was to administer tests attended a two hour training session at either the College of the Virgin Islands St. Thomas or St. Croix campuses. During this time the purpose of the testing was explained, the test and instruction manual were reviewed, a testing schedule was distributed and reviewed, and testing materials were distributed. These included a practice test for each of grades 2,4, and 6. This was to be given to students the day prior to the first day of testing in order to give them practice in reading and answering multiple choice standardized tests.

Tests were administered in the St. Thomas/St. John district during the week of October 21, 1980 and in the St. Croix district during the week of December 1, 1980. Testing materials and completed answer sheets were collected, answer sheets checked to determine compliance with marking instructions, and answer documents were sent to the Psychological Corporation of Iowa City, Iowa to be machine scored.

ERIC

Table 3

Stanford Subtests Administered at Each Grade Level

Grade	Subtest	Number of Items
Grade 12	Reading	78
(Task II Level	Mathematics	48
	English	69
Grade 10		
(Task I Level)	Reading	78
	Mathematics	48
	English	69
Grade 8	Vocabulary	50
(Advanced Level)	Reading Comprehension	74
	Mathematics Concepts	35
	Mathematics Computation	45
	Mathematics Applications	40
	Spelling	60
	Language	80
Grade 6	Vocabulary 50	
(Intermediate II Level)	Reading Comprehension	71
	Mathematics Concepts	35
	Mathematics Computation	45
	Mathematics Applications	40
	Spelling	60
	Word Study skills	50
	Language	80
Grade 4	Vocabulary	45
(Primary Level III)	Reading Comprehension	70
	Word Study Skills	55
	Mathematics Concepts	32
	Mathematics Computation	36
	Mathematics Applications	28
	Spelling	47
	Language	55
Grade 2	Vocabulary	37
(Primary Level I)	Reading Comprehension	87
	Word Study Skills	60
	Mathematics Concepts	32
	Mathematics Computation	32
	Listening Comprehension	26

Results

Table 4 provides descriptive statistics using the raw scores (number of items correct) of students on each subtest of the Stanford Achievement Test.

Content Validity

The content validity of the various levels of the Stanford Achievement Test used to collect data on basic skills achievement was determined by using the following strategies:

1) Collection of written curriculum guides used in the public schools. The objectives explicitly stated or implicitly inferred in these documents were compared with the lists of objectives tested provided by the test publisher.
2) Text books used in the teaching of basic skills subject matter were collected from selected schools. Stated and implicit objectives in these texts were compared with the test publisher's objectives.
3) The test objectives were shown to elementary and secondary subject area supervisors who were asked to determine the degree of match between those objectives and what is taught in the public schools at the indicated grade levels.
4) Selected building principals in St. Thomas were asked to review the objectives of the test and give their opinions concerning the degree of match between these objectives and the objectives taught toward in the classes in their schools.

Table 4

Descriptive Statistics of Stanford Achievement Test Raw Scores

Test	U.S.V.I. System Mean	U.S.V.I. System Stand. Dev.	STT/STJ Mean	STT/STJ Stand. Dev.	STX Mean	STX Stand Dev.
Grade 12-Task II Level						
Reading	43.9	13.8	40.6	12.3	48.3	15.2
Mathematics	25.3	8.3	24.6	7.4	26.3	9.4
English	46.8	11.2	45.6	10.9	48.4	11.5
Grade 10-Task I Level						
Reading	45.6	14.0	43.6	14.2	48.5	14.3
Mathematics	32.0	14.6	32.1	16.9	31.5	8.6
English	48.0	12.0	47.6	10.8	49.0	14.0
Grade 8-Advanced Level						
Vocabulary	21.0	7.2	20.7	6.6	21.2	8.5
Reading Comprehension	31.5	15.4	32.6	17.4	30.5	13.1
Mathematics Concepts	15.3	5.8	16.4	6.0	14.2	5.3
Mathematics Computation	23.0	7.6	23.3	7.1	22.8	8.2
Mathematics Application	16.9	6.6	17.7	6.6	16.1	6.5
Spelling	31.8	12.3	32.8	11.8	30.8	12.9
Language	35.6	12.1	35.8	10.6	34.6	13.6
Grade 6-Intermediate II Level						
Vocabulary	21.6	7.8	22.6	8.1	19.7	6.8
Reading Comprehension	32.5	12.4	31.8	12.7	33.5	11.5
Word Study Skills	28.6	11.1	29.4	11.5	27.1	10.3
Mathematics Concepts	18.2	5.7	19.3	5.5	16.4	5.5
Mathematics Computation	25.0	7.4	24.8	8.0	25.4	6.3
Mathematics Applications	18.4	8.0	19.2	8.0	16.9	7.8
Spelling	35.2	13.6	35.5	14.2	34.7	12.5
Language	37.4	13.8	37.9	15.0	37.0	11.5
Grade 4-Primary III Level						
Vocabulary	23.6	7.2	23.4	5.9	24.0	8.4
Reading Comprehension	42.3	11.5	42.2	11.3	42.3	11.8
Word Study Skills	29.8	10.0	30.8	9.2	28.7	10.8
Mathematics Concepts	15.5	5.3	15.0	4.4	16.0	6.2
Mathematics Computation	20.4	6.1	19.6	5.1	21.4	7.0
Mathematics Applications	13.8	5.8	13.8	5.5	13.8	6.0
Spelling	30.9	10.2	30.4	9.2	31.5	11.2
Language	28.9	8.8	28.1	8.1	29.8	9.4

Test	U.S.V.I. System Mean	Stand. Dev.	STT/STJ Mean	Stand. Dev.	STX Mean	Stand. Dev.
		Grade 2-Primary I Level				
Vocabulary	21.7	5.1	22.7	4.8	20.1	5.1
Reading (Part A)	34.1	13.8	36.3	15.2	30.5	10.3
Reading (Part B)	29.5	8.9	30.7	8.4	27.6	9.5
Word Study Skills	47.5	9.4	48.7	8.7	45.6	10.1
Mathematics Concepts	19.0	4.4	19.7	4.3	18.7	5.6
Mathematics Computation	21.5	5.0	22.0	4.6	20.7	5.4
Listening Comprehension	16.8	4.3	17.9	4.0	15.2	4.4

23

5) Teachers who administered the tests in their classrooms were asked to review the test publisher's objectives and to determine the degree of match between these objectives and the basic skills they expected their students to have obtained.

Using these techniques, the researchers were satisfied that the test did, indeed, test a sample of objectives that was consistent with the objectives used in teaching in the public schools of the Virgin Islands of the United States.

Reliability

The estimates of reliability of the test scores are presented in Table 5. The KR-20 reliability estimate[2] for each test is reported along with the KR-20 estimate for the mainland standardization samples as presented in the Technical Data Report. The issue of interpreting these reliability estimates is a complex one and will be dealt with in more detail at the conclusion of this report. The author felt the need to have at least a tentative criterion for making decisions regarding the acceptability of the reliability estimates obtained from the V.I. sample of examinees. The Stanford Achievement Test is considered

---

2

$$r_{xx} = [n/(n-1)]\,[\sigma^2_x - \Sigma pq/(\sigma^2_x)]$$

where r_{xx} = the reliability estimate
n = the number of scores
σ^2_x = the variance of the distribution of scores
p = the proportion of examinees marking the correct answer on a particular item
q = 1-p

(From Stanford Achievement Test: Technical Data Report, p. 35) to have more than acceptable reliability when administered to the population of examinees upon which it was standardized (i.e. continental U.S. students). Among the indications of this are numerous reviews of the test in the literature (Kasdon, 1974; Lehmann, 1975; Chase, 1978; Ebel, 1978; Thorndike, 1978) and the fact that it is widely used in the schools. However, the literature is replete with studies which indicate that standardized tests of academic achievement tend to produce less reliable scores when administered to students from low socioeconomic status homes and to those who are culturally different from the majority of those on whom the test was normed (see reviews and discussions in Anastasi, 1958; Tyler, 1956; and Deutsch, 1960). Therefore, if the reliability estimates obtained from a sample of U.S. Virgin Islands students who took the Stanford Achievement Test are at least equal to the reliability estimates obtained from the standardization samples, it is reasonable to conclude that the test scores are reliable indicators of academic achievement for these students.

For each reliability estimate obtained from the V.I. sample, a reliability difference was found by subtracting the standardization groups' reliability estimate from the local groups' reliability estimates. The distribution of these differences is shown by the histogram in Figure 1. The median reliability difference was -.038 with a range from -.20 to +.05 with the distribution skewed to the left (i.e. negatively) quite markedly.

In addition in an attempt to observe these reliability differences from another perspective, for each pair of reliability estimates (the standardization group estimate and the

Table 5

Stanford Achievement Test Raw Score Reliability Estimates

TEST	STAND. GROUPS KR-20	USVI SYSTEM KR-20	ST THOMAS/ ST JOHN KR-20	ST CROIX KR-20
Grade 12-TASK II Level				
Reading	.94	.93	.91*	.95
Mathematics	.94	.91*	.90*	.91
English	.94	.87*	.84*	.90
Grade 10-TASK I Level				
Reading	.95	.93*	.94	.94
Mathematics	.94	.98	.99	.89*
English	.95	.92*	.90*	.95
Grade 8-Advanced Level				
Vocabulary	.89	.81*	.78*	.87
Reading Conprehension	.94	.95	.97	.93
Mathematics Concepts	.86	.74*	.81*	.75*
Mathematics Computation	.89	.85*	.83*	.87
Mathematics Application	.91	.83*	.83*	.82*
Spelling	.94	.93	.92*	.93
Language	.94	.88*	.84*	.91*
Grade 6-Intermediate II Level				
Vocabulary	.90	.85*	.86*	.80*
Reading Comprehension	.94	.93	.94	.91
Word Study Skills	.95	.93*	.94	.92*
Mathematics Concepts	.85	.79*	.78*	.77*
Mathematics Computation	.90	.85*	.88	.78*
Mathematics Application	.92	.89*	.89*	.89
Spelling	.94	.94	.95	.93
Language	.94	.92*	.93	.87
Grade 4-Primary III Level				
Vocabulary	.88	.83*	.75*	.88
Reading	.96	.91*	.91*	.92*
Word Study Skills	.94	.90*	.88*	.92
Mathematics Concepts	.86	.77*	.66*	.84
Mathematics Computation	.87	.83*	.75*	.87
Mathematics Application	.92	.86*	.84*	.87*
Spelling	.93	.93	.91*	.94
Language	.92	.86*	.84*	.88*

Table 5 (cont.)

TEST	STAND. GROUPS KR-20	USVI SYSTEM KR-20	ST THOMAS/ ST JOHN KR-20	ST CROIX KR-20
	Grade 2-Primary I Level			
Vocabulary	.86	.72*	.71*	.71*
Reading Part A	.94	.98	.99	.94
Reading Part B	.95	.92*	.91	.92*
Word Study Skills	.93	.92	.91	.93
Mathematics Concepts	.81	.71*	.69*	.74*
Mathematics Computation	.87	.81	.78*	.83*
Listening Comprehension	.77	.74	.70*	.72

*Significantly lower than the standardization groups KR-20 at p=.05

32

V.I. sample estimate), the hypothesis that the differences in reliability obtained were less than zero was tested. Reliability estimates were transformed using Z transformations to normalize the skewness of the distribution of the correlation measures and the hypothesis tested with t-tests[3]. One tailed significance tests were used. As indicated by Table 5, 64 of 108 comparisons show lower reliability in the V.I. sample (i.e. differences less than zero).

A note of caution is in order in interpreting the results of these tests of significant differences. As previously pointed out, cluster sampling was used in obtaining the sample of Virgin Island students to be tested rather than simple random sampling. The result of this is that the actual standard error of the sample may very well be larger than the one used in calculating the t statistic (estimated by $[1/(N-3)]^{\frac{1}{2}}$). The result of this would be that the values of t obtained were larger than they should have been and that some of the differences from zero that were noted in Table 5 to be significant at the p=.05 level may actually not have been. To put it in technical terms, the probability of Type I error is probably greater than .05 in each of these hypothesis tests. This is a definite weakness in any conclusions we might draw from these tests. However, from a practical point of view, given the decisions to be made,

---

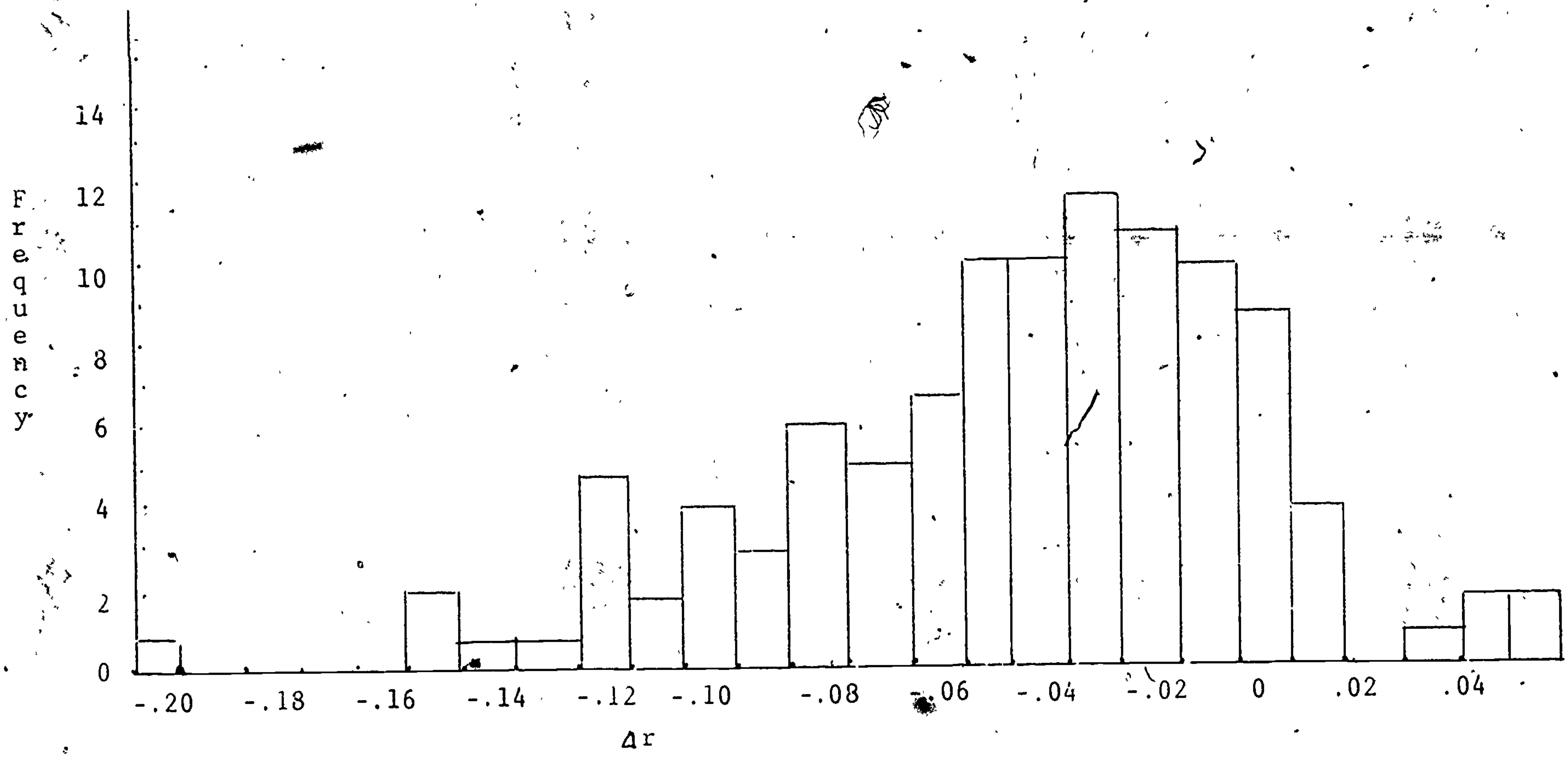
3

$$Z = \tfrac{1}{2} \log_e \frac{(1+r_{xy})}{(1-r_{xy})}$$ (Hayes, 1973, pp. 662-667)

$$t = \frac{Z - E(Z)}{\sqrt{1/(N-3)}}$$

Figure 1

Frequency Distribution of Differences Between the Standardization Group Reliability Estimates and the V.I. Sample Reliability Estimates (Δr)



Type I error is the error of preference. That is, the consequences of mistakenly assuming that scores are less reliable for V.I. would be that we would either look more closely at these tests from which the scores came or discard the results of the testing as being unreliable for V.I. students. In this case what is lost is much time and, possibly, some money. On the other hand, the consequence of Type II errors (mistakenly assuming that local scores are at least as reliable as the standardization groups' scores) would be to go ahead and use the unreliable scores to make decisions about basic skills levels of V.I. students and, possibly, to make decisions regarding instructional strategies that will be used in the schools. In essence, then, what results is a rather liberal test of the hypotheses and, given the nature of the decisions to be made, this may not be totally undesirable. However, it must be kept in mind when interpreting these results that the actual level of Type I error is <u>not</u> known and that it is probably higher than .05. In any event, we can use Table 5 to flag tests where reliabilities <u>may</u> be less than acceptable.

It was noted that, in the majority of cases, the variances of the raw scores obtained by the V.I. sample were considerably lower than those reported for the standardization groups. This homogeneity is a phenomenon commonly found when testing samples drawn from populations composed largely of persons from low socioeconomic status homes. "The reliability of any test is partially dependent on the sample of individuals tested to obtain the coefficient. In general, the more heterogeneous the sample with respect to whatever the test is measuring, the higher the reliability coefficient will be" (Technical Data Report), p. 35). The standardization groups' reliability coefficients can be adjusted for homogeneity using the variances obtained from the local sample making reliability comparisons more meaningful.[4]

Using the adjusted reliability estimates for the standardization groups, the differences between this group's and the V.I. sample's reliability estimates was calculated employing the same procedure used with the unadjusted estimates. The distribution of these differences is shown in the histogram in Figure 2. The median reliability difference using the adjusted estimates was -.002 with a range from -.06 to .02. As with the unadjusted scores, the distribution is negatively skewed, but not as markedly as with the unadjusted reliability estimate differences. When the standardization group's reliability estimates are adjusted for homogeneity, the differences between the reliability estimates of the two samples become fewer and smaller.

Table 6 presents the adjusted estimates of reliability for the standardization groups' and the results of tests of the hypotheses that the differences between the standardization groups' reliability estimates and the estimates of reliability

---

4 Using the formula $\rho_{xy} = (1-\sigma^2_{x*}/\sigma^2_x)(1-\rho_{x*y*})$

where ρ_{xy} and σ^2_x are, in this case, the reliability coefficient and variance of the standardization groups and ρ_{x*y*} and σ^2_{x*} are the same statistics for the V.I. sample.

(Hayes, 1973)

Figure 2

Frequency Distribution of Differences Between the Standardization Group Adjusted Reliability Estimates and the V.I. Sample Reliability Estimates (Δr)

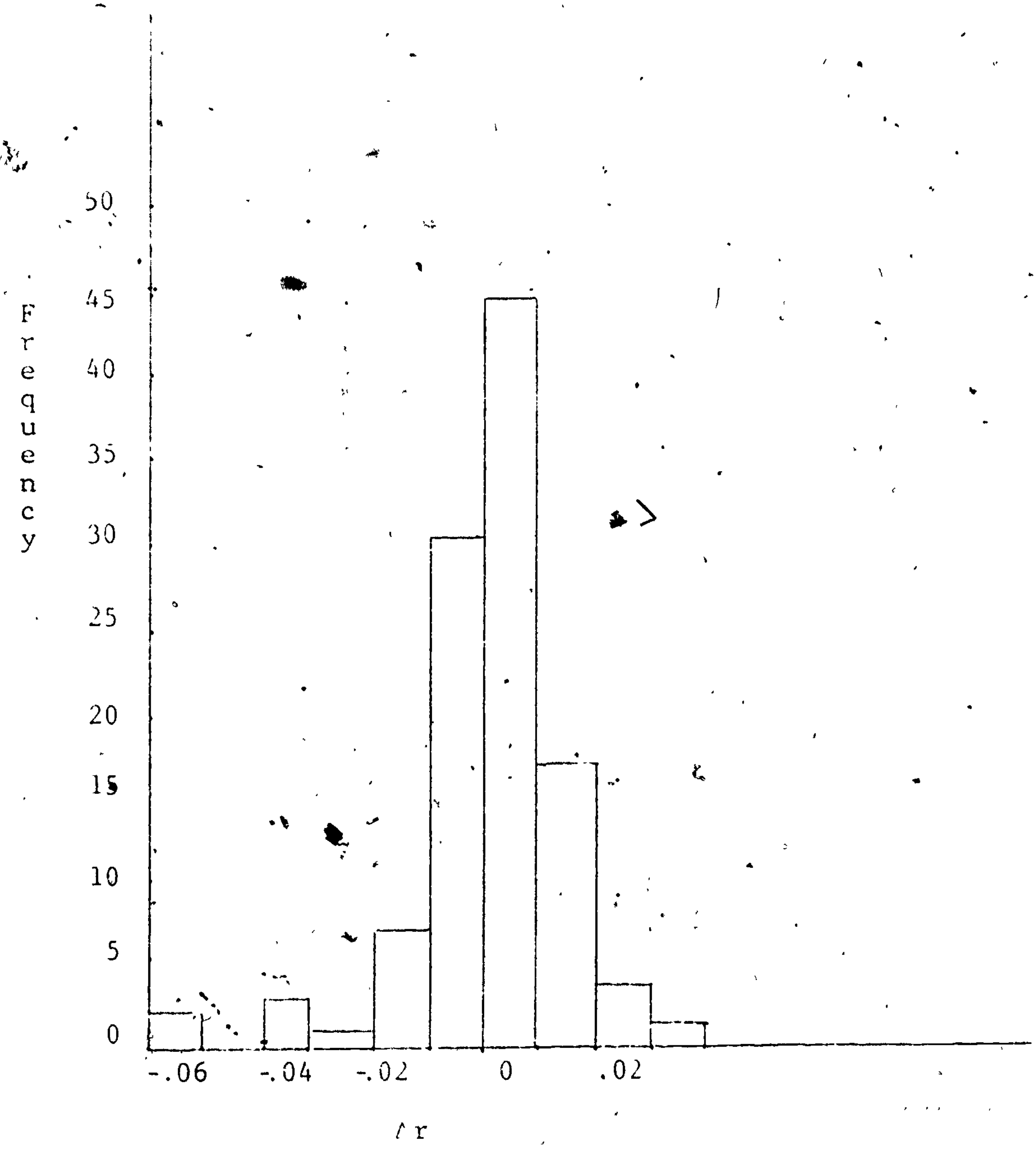


Table 6

Adjusted Stanford Achievement Test Raw Score Reliability Estimates

TEST	USVI SYSTEM		ST THOMAS/ST JOHN		ST. CROIX	
	Adj. Stand. Groups KR-20	Local Sample KR-20	Adj. Stand. Groups KR-20	Local Sample KR-20	Adj. Stand. Groups KR-20	Local Sample KR-20
Grade 12 - TASK II Level						
Reading	.93	.93	.91	.91	.94	.95
Mathematics	.88	.87	.85	.84	.91	.90
English	.91	.91	.91	.90	.91	.91
Grade 10 - TASK I Level						
Reading	.93	.93	.94	.94	.94	.94
Mathematics	.97	.98	.97	.99	.90	.89
English	.93	.92	.92	.90	.95	.95

ERIC

Table 6 (cont.)

TEST	USVI SYSTEM		ST THOMAS/ST JOHN		ST. CROIX	
	Adj. Stand. Groups KR-20	Local Sample KR-20	Adj. Stand. Groups KR-20	Local Sample KR-20	Adj. Stand. Groups KR-20	Local Sample KR-20
Grade 8 - Advanced Level						
Vocabulary	.82	.81	.79	.78	.87	.87
Reading Comprèsension	.94	.95	.95	.97	.92	.93
Mathématics Concepts	.79	.79	.80	.81	.75	.75
Mathematics Computation	.85	.85	.83	.83	.87	.87
Mathematics Application	.83	.83	.83	.83	.83	.82
Spelling	.93	.93	.92	.92	.93	.93
Language	.90	.88*	.86	.84	.92	.91
Grade 6 - Intermediate II Level						
Vocabulary	.86	.86	.87	.86	.81	.80
Reading Comprehension	.92	.93	.93	.94	.91	.91
Work Study Skills	.93	.93	.94	.94	.92	.92
Mathematics Concepts	.79	.79	.78	.78	.77	.77
Mathematics Computation	.85	.85	.87	.88	.79	.78
Mathematics Application	.90	.89	.90	.89	.89	.89
Spelling	.94	.94	.95	.95	.93	.93
Language	.92	.92	.92	.93	.88	[illegible]

Table 6 (cont.)

TEST	USVI SYSTEM Adj. Stand. Groups KR-20	USVI SYSTEM Local Sample KR-20	ST THOMAS/ST JOHN Adj. Stand. Groups KR-20	ST THOMAS/ST JOHN Local Sample KR-20	ST. CROIX Adj. Stand. Groups KR-20	ST. CROIX Local Sample KR-20
Grade 4 - Primary III						
Vocabulary	.84	.83	.76	.75	.88	.88
Reading Comprehension	.93	.91*	.92	.91	.93	.92
Word Study Skills	.91	.90	.89	.88	.92	.92
Mathematics Concepts	.78	.77	.68	.66	.84	.84
Mathematics Computation	.83	.83	.76	.75	.87	.87
Mathematics Application	.87	.86	.85	.84	.88	.87
Spelling	.93	.93	.91	.91	.94	.94
Language	.86	.86	.84	.84	.88	.88
Grade 2 - Primary I Level						
Vocabulary	.76	.72	.72	.71	.76	.72
Reading - Part A	.97	.98	.97	.99	.94	.94
Reading - Part B	.93	.92	.92	.91	.94	.92
Work Study Skills	.91	.92	.89	.91	.92	.93
Mathematics Concepts	.72	.71	.70	.69	.74	.74
Mathematics Computation	.80	.81	.76	.78	.82	.83
Listening Computation	.78	.74	.73	.70	.78	.72

*Significantly lower than the standardization groups' KR-20 at p=.05

ERIC

for the scores obtained by the V.I. sample are less than zero. Again, the caution mentioned in discussing the hypothesis tests using the unadjusted reliability estimates holds true. The actual level of Type I error involved in these tests is not really known and is most probably higher than .05. Even under this condition, however, only 2 out of 108 comparisons showed differences significantly less than zero. Since the p=.05 level was formally used, these two differences could be expected as a result of Type I error (i.e. as a result of chance). In fact, slightly more than 5 differences significantly less than zero would have been expected on a chance basis.

The standard error of measurement[5] for the raw scores of the V.I. sample are shown in Table 7. ". . .when the reliability of a test is interpreted in terms of the standard error of measurement, the problem of the influence of heterogeneity [or homogeneity] is taken into account, since the formula for the standard error of measurement includes the standard deviation of the scores" (Technical Data Report, p. 35). The standard error or measurement can be thought of as the standard deviation of the differences between the scores obtained on the test and the true scores (the scores the examinees would have received if the test were perfectly reliable). As such, it can be used to determine an interval within which we can be confident that the true score falls. For instance, we can be confident that the true score would be within one standard

---

5

$SE = S\sqrt{1-r_{xx}}$ where SE is the standard error of measurement S is the standard deviation of the scores, and r_{xx} is the reliability coefficient.

(Gronlund, 1976)

Table 7

Stanford Achievement Test
Standard Error of Measurement Estimates

TEST	STAND. GROUPS S.E.M.	USVI SYSTEM S.E.M.	ST THOMAS/ ST JOHN S.E.M.	ST CROIX S.E.M.
Grade 12 - TASK II Level				
Reading	2.60	3.65*	3.69*	3.40*
Mathematics	2.80	3.01	2.98	2.97
English	3.30	3.36	3.45	3.44
Grade 10 - TASK I Level				
Reading	2.50	3.70*	3.48*	3.50*
Mathematics	2.60	2.07	1.69	[illegible].85
English	3.10	3.38*	3.41*	3.70*
Grade 8 - Advanced Level				
Vocabulary	3.10	3.15	3.10	3.05
Reading Comprehension	3.60	3.45	3.02	3.46
Mathematics Concepts	2.60	2.65	2.61	2.67
Mathematics Computation	2.90	2.95	2.92	2.94
Mathematics Application	2.60	2.72	2.72	2.76
Spelling	3.30	3.27	3.33	3.40
Language	3.90	4.20*	4.23	4.07

Table 7 (con't.)

TEST	STAND. GROUPS S.E.M.	USVI SYSTEM S.E.M.	ST THOMAS/ ST JOHN S.E.M.	ST CROIX S.E.M.
Grade 6 - Intermediate II Level				
Vocabulary	3.0	3.02	3.05	3.05
Reading Comprehension	3.50	3.29	3.11	3.46
Word Study Skills	2.90	2.93	2.82	2.93
Mathematics Concepts	2.60	2.60	2.59	2.64
Mathematics Computation	2.90	2.88	2.78	2.94
Mathematics Application	2.60	2.63	2.64	2.60
Spelling	3.30	3.33	3.18	3.30
Language	4.0	3.91	3.98	4.14
Grade 4 - Primary III Level				
Vocabulary	2.80	2.96	2.95	2.91
Reading Comprehension	3.20	3.46*	3.38	3.35
Word Study Skills	3.00	3.16	3.17	3.06
Mathematics Concepts	2.40	2.55	2.55	2.47
Mathematics Computation	2.50	2.52	2.55	2.47
Mathematics Application	2.0	2.16*	2.21*	2.18
Spelling	2.70	2.69	2.76	2.74
Language	3.10	3.28*	3.26	3.25

47

Table 7 (cont.)

TEST	STAND. GROUPS S.E.M.	USVI SYSTEM S.E.M.	ST THOMAS/ ST JOHN S.E.M.	ST CROIX S.E.M.
Grade 2 - Primary I Level				
Vocabulary	2.50	2.68	2.57	2.72
Reading - Part A	2.50	1.95	1.52	2.51
Reading - Part B	2.40	2.53	2.52	2.68
Word Study Skills	2.80	2.65	2.62	2.68
Mathematics Concepts	2.30	2.38	2.38	2.37
Mathematics Computation	2.20	2.17	2.15	2.21
Listening Comprehension	2.0[illegible]	2.22*	2.17	2.34*

*Significantly higher than the Standardization Groups at the p=.05 level.

43

ERIC Full Text Provided by ERIC

error of the score the student actually received (the observed score) around 68% of the time. The true score would be within two standard errors of the observed score approximately 96% of the time. Naturally, the lower the standard error of measurement, the more reliable the scores.

χ^2 (chi-squared) tests[6] were used to test the hypotheses that the standard errors of measurement for the test scores in the Virgin Islands sample were greater than those for the standardization sample. Sixteen of the 108 tests show significantly higher standard errors for the V.I. sample at the p=.05 level of significance. Nine of them occur in the high school tests in reading and English areas. These tests need to be looked at closely. Among the remaining seven there seem to be no patterns. It should be noted, however, that in two of these cases, Mathematics Applications in grade 4 and Listening Comprehension in grade 2, the differences are in one district and in the total system scores. Since the total system scores are obviously affected by the individual district scores, it is possible that the large total system standard errors may be a result of the lower reliability obtained from the district scores.

6

$\chi^2/df = S^2/\sigma^2$

where df is the number of degrees of freedom,
S^2 is the square of the V.I. sample standard error of measurement,
σ^2 is the square of the standardization groups standard error of measurement.
(Darlington, 1975)

Summary and Conclusions

The scores obtained from the testing of a representative sample of U.S. Virgin Islands students using the 1973 edition of the Stanford Achievement Test appear to be both content valid and reliable. This is significant in that this test, and all standardized tests of academic achievement published in the United States, have been designed without including studies of noncontinental U.S. public school curriculum in the test planning process or using noncontinental U.S. students in its standardization studies.

It is clear that the test objectives, as stated by the publisher, are a good match for those used in U.S. Virgin Islands public schools. In addition, the reliability estimates of the scores obtained from Virgin Islands students are, in most cases, not significantly different from those obtained using the continental U.S. standardization samples. At this point it may be useful to examine the distinction between differences that are "statistically significant" and those that are "educationally significant." The statement that two values are "statistically significant" implies that we are confident that the difference between the two values <u>is not zero</u>. This is no guarantee that the differences <u>are not trivial</u>. For instance, we may weigh two packages on the same, very accurate, scale and find that one weighs 25 kilograms while the other weighs 25.5 kilograms. If we were trying to decide which of these packages to assign each of two people to carry based on their relative strengths, we could probably conclude that either person could carry either package. The difference of one half of a kilogram was real (i.e. nonzero), but it was so small that it was trivial. Likewise, differences in reliability estimates noted in this study may be statistically significant, but so small as to allow us to conclude that the test scores were reliable enough for us to use to make educational decisions (i.e. the differences were not educationally significant). With the exception of the grade 12 and grade 10 Reading test scores and the grade 10 English test scores from the St. Croix district, the differences observed in standard error of measurement estimates seem to be so small as not to be educationally significant.

Putting aside the question of the comparability of the obtained reliability estimates between the standardization samples and the U.S. Virgin Islands sample, the question of whether or not the scores obtained from the U.S.V.I. sample are reliable enough for us to use them to make educational decisions needs to be addressed. "The degree of reliability we demand in our educational measures depends largely on the nature of the decision to be made" (Gronlund, 1976, p.124). Standardized test results are used by school personnel as one source of information for making instructional and curricular decisions. Other sources of information such as teacher made classroom tests and observational techniques are combined with the results of standardized tests before final educational decisions are made in schools. Finally, this particular study was designed to point out strengths and weaknesses in basic skills areas in U.S. Virgin Islands public schools. Those

ERIC

persons entrusted with the responsibility for making curricular and instructional decisions in the Department of Education will use this and other information before making changes in what goes on in schools. Further, decisions made will always be open to confirmation and change. Cronbach (1970) points out that the reversability of decisions made on the basis of test data is an important factor to take into consideration in making judgements concerning desired levels of reliability. The reliability estimates obtained from the U.S.V.I. sample which seem to cluster from the middle .80's to the middle .90's are more than adequate to allow the confident use of the obtained scores.

References

Anastasi, A. Differential Psychology (3rd ed.). New York: Macmillan, 1958.

Asher, W.J. Educational Research and Evaluation Methods. Boston: Little, Brown & Co., 1976.

Chase, C.I. Review of the Test of Academic Skills. In O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook (vol. 1). Highland Park, N.J.: Gryphon Press, 1978.

Cronbach, L.J. Essentials of Psychological Testing (3rd ed.). New York: Harper and Row, 1970.

Cronbach, L.J. & Meehl, P.E. Construct validity in psychological tests. Psychological Bulletin, 1955, 52, 281-302.

Darlington, R.B. Radical and Squares. Ithaca, N.Y.: Logan Hill Press, 1975.

Deutsch, M. Minority group and class status as related to social and personality factors in scholastic achievement (Monograph No. 2). Ithaca, N.Y.: The Society for Applied Anthropology, 1960.

Ebel, R.L. Must all tests be valid? American Psychologist, 1961, 16, 640-643.

Ebel, R.L. Review of the 1973 edition of the Stanford Achievement Test. In O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook (vol.1). Highland Park, N.J.: Gryphon Press, 1978.

Ebel, R.L. Essentials of Educational Measurements (3rd ed.). Englewood Cliffs, N.J.: Prentice-Hall, 1979.

French, J.W. & Michael, W.B. (Cochairmen) Standards for Educational and Psychological Tests and Manuals. Washington, D.C.: American Psychological Association, 1966.

Gronlund, N.E. Measurement and Evaluation in Teaching (3rd ed.). New York: Macmillan, 1976.

Hayes, W.L. Statistics for the Social Sciences (2nd ed.). New York: Holt, Rinehart, and Winston, 1973.

Kasdon, L.M. The Stanford Achievement Test (Review of the 1973 edition of the Stanford Achievement Test). Reading Teacher, 1974, 27, 743+.

Lehmann, I.J. The Stanford Achievement Test Series - 1973 (Review of the 1973 edition of the Stanford Achievement Test). Journal of Educational Measurement, 1975, 12, 297-306.

Passow, A.H. Review of the 1973 edition of the Stanford Achievement Test. In O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook (vol. 1). Highland Park, N.J.: Gryphon Press, 1978.

Stanford Achievement Test. Technical Data Report. New York: Harcourt Brace Jovanovich, 1975.

Thorndike, R.L. Review of the Test of Academic Skills. In O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook (vol. 1). Highland Park, N.J.: Gryphon Press, 1978.

Tyler, L. The Psychology of Individual Differences (2nd ed.). New York: Appleton-Century-Crofts, 1956.

Warwick, D.P. & Lininger, C.A. The Sample Survey: Theory and Practice. New York: McGraw-Hill, 1975.