DOCUMENT RESUME

ED 092 105	40	IR 000 664

AUTHOR	Bond, Jack H.
TITLE	Verifying Sample Size Concerns. Research Report No. 737.
INSTITUTION	Syracuse City School District, N.Y.
SPONS AGENCY	Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C. Media Services and Captioned Films Branch.
PUB DATE	Apr 73
CONTRACT	OEC-9-423617-4357(616)
NOTE	12p.; Supporting document for the Computer Based Project for the Evaluation of Media for the Handicapped
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	*Educable Mentally Handicapped; Evaluation; *Exceptional Child Research; Intermediate Grades; *Media Research; *Predictive Validity; Primary Grades; *Sampling; Statistical Studies; Student Testing; Testing
IDENTIFIERS	*Computer Based Project Evaluation Media Handicap

ABSTRACT

A study was made to ascertain how large a sample is needed to make media effectiveness decisions which are generalizable to the total educable mentally handicapped (EMH) population. The method employed in the study involved pretesting and posttesting a sample of 70 primary and intermediate EMH children on the content of a filmstrip. Statistical analyses of the data indicated that samples of five students gave results that were within the parameters of decision established by the Computer Based Project (Syracuse, N.Y.). When the sample size was increased to 10, the standard findings for increased sample size were supported, i.e., scores were within smaller ranges, variance between groups was reduced, and gains were more standardized. However, the investigators concluded that samples of five subjects seem to be large enough to establish estimates of population parameters within the limits of four out of five times. Larger samples do not add appreciable data or substantially change the outcome of decisions obtained from the samples of five. (Author/WCM)

Research Report #737
Jack H. Bond
April 1973

VERIFYING SAMPLE SIZE CONCERNS

ABSTRACT

The purpose of this study was to ascertain how large a sample is needed to make media effectiveness decisions which are generalizable to the total educable mentally handicapped (EMH) population. The literature regarding sample size consideration was reviewed. The method employed in this study involved pretesting and posttesting a sample of seventy primary and intermediate EMH children on the content of a filmstrip. Statistical analyses of the data indicated that samples of five students gave results that were within the parameters of decision established by the Computer Based Project (Syracuse, N. Y.) . When the sample size was increased to ten, the standard findings for increased sample size were supported, i. e. scores were within smaller ranges, variance between groups was reduced and gains were more standardized. However, the investigators concluded that samples of five subjects seem to be large enough to establish estimates of population parameters within the limits of four out of five times. Larger samples do not add appreciable data or substantially change the outcome of decisions obtained from the samples of five.

SPECIAL REPORT No. 737

COMPUTER-BASED PROJECT for the EVALUATION of MEDIA for the HANDICAPPED

Title: VERIFYING SAMPLE SIZE CONCERNS

BACKGROUND

The Computer Based Project for the Evaluation of Media for the Handicapped, based on contract #OEC-9-423617-4357 (616) between the Syracuse (N.Y.) City School District and the Media Services and Captioned Films Branch, Bureau of Education for the Handicapped (United States Office of Education) for the five year period July 1, 1969 through June 30, 1974. The major goal is to improve the instruction of handicapped children through the development and use of an evaluation system to measure the instructional effectiveness of films and other materials with educable mentally handicapped (EMH) children, in-service training and media support for special teachers, and studies related to the evaluation process and the populations used.

The Project has concentrated on the 600 films and 200 filmstrips from the Media Services and Captioned Films (BEH - USOE) depository; however, specific packages from Project LIFE, various elementary math curricula, and selected programs from Children's TV Workshop have also been evaluated. The evaluation model used requires that: 1) objectives of materials be specified and written; 2) instruments be constructed to test and measure effectiveness; and, 3) children be the major sources of evaluation information. A number of instruments and methodologies are employed in the gathering of cognitive and affective data from 900 EMH children and 80 special teachers to make the effectiveness decisions. Over half of the EMH population can neither read or write; therefore, a unique Student Response System (SRS) is employed, consisting of a twenty station G.E.-1000 SRS which can be operated in a group or individual recording mode and is connected to a remote computer system. The computer capabilities consist of remote telephone connections to the Rome (N.Y.) Air Development Command, the Honeywell time-shared network, and the Schenectady (N.Y.) G E Research and Development Center; and batch mode capabilities of the Syracuse City Schools, Syracuse University, and various commercial sources.

In-service and media support activities provide on-the-job training for teachers, teacher aides, equipment, and materials to the special teachers in the city schools. The research activities have centered around investigations and special problems related to the development of the evaluation model. The four major areas considered are: 1) testing effects, 2) captioning effects, 3) special student characteristics; and, 4) evaluation procedures validation.

Documentation of the major activities appear in the five annual reports and the 600 evaluations prepared on materials used. Staff members were encouraged to prepare special reports and the attached paper is one of these. The opinions expressed in this publication do not necessarily reflect the position or policy of the Computer Based Project, the United States Office of Education, or the Syracuse City School District, and no official endorsement by any of the agencies should be inferred.

Research Report #737
Jack H. Bond
April 1973

VERIFYING SAMPLE SIZE CONCERNS

The problem has arisen as to how large a sample is needed to make media effectiveness decisions generalizable to the total EMH population.

In an effort to obtain the best possible effectiveness decision using the least number subjects and replications, this study investigated sample size considerations. This kind of research has major implications for cost-effectiveness; if fewer subjects are needed to make the decisions, more evaluations can be run on the same time parameters using regular Project personnel.

BACKGROUND

Decisions are currently being made based on a sample of five or more subjects who make a score of 50% or better correct on a cognitive measure subsequent to viewing a film or filmstrip; or on subjects who make a positive gain in a pretest-posttest design. In the latter case, an arbitrary 20% gain is used as a criterion, attenuating it by subjective considerations, where the percentage was close to the criterion. Some concern and question has arisen as to the efficacy, reliability, and validity of making decisions on such small sample sizes.

The literature suggests several similar concerns as indicated in the following comments.

For standard parametric statistical procedures, Hays (1966) suggests that, in experiments where there is little "natural" variation in materials observed that small samples can give the same power as large samples in those experiments which have extremely variable conditions. A number of authors have been more pragmatic and practical, they simply suggest rather strongly that samples under thirty be avoided, if possible, when using parametric statistics (Gilford, 1965; Winer, 1954; Cohen,1969). Non-parametric statistics are more applicable to samples as small as N = 3, but suggest that groups between 3 and 30 give more reliable results (Seigel, 1956). Chi square concerns suggest that for degrees of freedom less than 30 that the Table of Chi Square be used, but for over 30 the distribution approximates the normal distribution (Linqinst, 1956, page 28) and that cell frequencies less than five in over twenty percent of the cells are to be avoided (Seigel, 1956, page 178). The relationship of sample size to reliability has been extensively discussed and pointed out. For example, Cohen relates that "reliability is always dependent upon sample size" (1969, page 6). He further states "The greater the degree to which the means of different samples vary among themselves, the less any of them can be relied upon (ibid,page 7)." Generally, an increase in sample size will reduce errors of measurement and thus increase reliability and the power of the test used (Cohen, 1969, page 11, 13; Seigel, 1956). The general procedure has been then to obtain as large a sample as possible and administer most of the items to all individuals in the sample. This usually is an arduous task for both the experimentor and the subjects, if done on a continuing basis.

The use of matrix sampling techniques suggests that it is unnecessary for all subjects to take all items and that just as valid population estimates can be obtained from sub-samples of items given to sub-samples of the population. Immediate concern is raised then for the best number of items in the sub-sample of items given to the best number of subjects. May and Barcikowski (1973) and Shoemaker (1971) recommended the fewer items (c.6 items) and several subtests made by single exhaustion of the item set give the best estimates of means and standard deviations (when biserial correlations are high, i.e., .45 to .95; when biserial correlation is not high, i.e., .05 to .35, they recommend larger subtests up to half the item population).

The question for this investigation then becomes how large a sample of respondents should be obtained to have estimates be stable in four out of five samples. That is, the probability of making a Type I error is equal to .20 (page 280, Hays, 1966).

METHOD

Several filmstrips in the CBP (Bond, 1972) Evaluation System have been shown to a number of children in which they were pretested before seeing the filmstrip and posttested after seeing it. A set of data for the filmstrip "Our Hands" was used for this study.

Instrument: The pretest and posttest percent correct responses from seventy primary and intermediate EMR children who were tested with a nine item multiple choice instrument before viewing a 24 frame filmstrip "Our Hands." They were again tested with the instrument after viewing. The percent of correct answers was computed for each student.

Sample: From the population of 70 EMH children at primary and intermediate level who responded to items prepared for the filmstrip "Our Hands" those scoring 80% to 100% on the pretest were dropped from the population. Five random samples of five subjects were selected by replacement after each sample from the remaining available population of 60.

The five sample parameters are summarized in Table 1. Five random samples of ten subjects were selected by replacement after each sample from the available population and sample parameters computed as shown in Table II.

Treatment: The pretest and posttest scores for the selected subject(S_s) were recorded from the line of student response data and included as a measure for the sample. The descriptive statistics - pretest mean, 5d, posttest mean and sd and difference (gain) are computed for each sample group of 5 and 10 S_s. These statistics were also computed for the total population of 60.

A pretest-treatment-posttest model served to design this study. The null hypothesis of no differences between all groups was tested in a ten-group-repeated-measures one way analysis of variance (ANOVA) is used for the sample group of 5 and the sample group of 10. A Spearman rank order correlation (rho) is computed for samples of 5 and 10 to indicate the correlation between the pretest/posttest scores. Test-sample interactions were tested using a two by five two-way ANOVA was used for the five samples in each to check for.

Analysis: The parameters of the total population of 60 were computed and are shown in both Table I and Table II below. A one-way analysis of variance was performed on the sample size of 5 data (and the sample of 10), in a ten-group-repeated design, testing the null hypothesis of no differences between groups. For the 5-student samples, the obtained F = 1.81 was not significant at the p = .05 level when a critical value of $F_{9,40}$ = 2.124 was necessary to reject the no difference. At the .80 confidence level, the differences between sample means needs to be 21.6 units or greater. The summary results are shown in Table I below.

TABLE I SUMMARY:

SAMPLE OF 5 STUDENTS

SAMPLE	SIZE	MEAN		STANDARD DEVIATION	%	SIGNIFICANCE*		
		Pretest	Posttest	at .95 Confolevel	Diff.	P=.05	P=.20	rho
A	5	59.7		19.32				
			77.5	10.07	17.8	n/s	n/s	-.07
B	5	33.2		18.41				
			62.1	11.46	28.9	n/s	s	-.10
C	5	42.0		19.07				
			62.0	31.98	20.0	n/s	n/s	.70
D	5	44.6		38.81				
			53.2	35.53	8.8	n/s	n/s	.90
E	5	32.2		17.43				
			71.6	14.72	37.8	s	s	.53
Sample population (N=25)								
		42.54	65.21	22.67				.17
Total Population (N=60)								
		47.03	62.04	15.01				

*At a confidence level P = .20, differences - 21.6
P = .05, differences - 32.6

For the 10 students samples, the obtain F = 1.27 was not significant at the P = .05 level when a critical value of $F_{9,40}$ = 2.124 was necessary to reject the null hypothesis. The summary results are shown in Table II.

TABLE II

SUMMARY: SAMPLE OF 10 SUBJECTS

SAMPLE	SIZE	MEAN		STANDARD DEVIATION	%	SIGNIFICANCE rho	
		Pretest	Posttest	at .95 level	Gain	P=.05	P=.20
1	10	36.3		15.50	17.0	NO	YES
			53.3	23.26			
2	10	42.2		19.94	14.5	NO	YES
			56.7	24.66			
3	10	40.7		22.02	11.2	NO	NO
			51.9	22.16			
4	10	44.2		15.70	10.0	NO	NO
			54.2	17.57			
5	10	35.4		23.7	16.8	NO	YES
			52.2	23.3			
SAMPLE GRAND MEAN (N=50)							
		39.84	53.60		13.76		
POPULATION GRAND MEAN (N=60)							
		43.16	61.88		18.72		

At Confidence Level P = .05, differences must be greater than 19.56
P = .20, differences must be greater than 12.76

The two data sets were regrouped into a 2 x 5 two-way ANOVA to check for interaction effects and to assume independence of pretest from posttest data. The results are shown in Table III for the 5 student data and in Table IV for the 10 student data.

TABLE III

SUMMARY: TWO-WAY ANOVA

5 STUDENT SAMPLES

SOURCE	SS	DF	M-SQUARE	F	df	SIGNIFICANCE*
TEST	6498.0	1	6498.0	9.886	1,40	P = .005
GROUPS	2895.88	4	723.97	1.101	4,40	NS
TEST X GROUPS	1297.00	4	324.25	.493	4,40	
SSE	26290.81	40	657.27			
SST	36981.69	49	754.73			

*Tabled $F_{1,40}$ = 8.8278 at P = .005 (Owen, 1962)

$F_{4,40}$ = 2.606 at P = .05

TABLE IV

SUMMARY: TWO-WAY ANOVA

10 STUDENT SAMPLES

SOURCE	SS	DF	M-SQUARE	F	df	SIGNIFICANCE*
TESTS	4830.25	1	4830.25	9.9219	1,90	P = .005
GROUPS	519.84	4	129.96	.2669	4,90	NS
TEST X GROUP	204.40	4	51.10	.1049	4,90	
SSE	43814.11	90	486.82			
SST	49368.60	99	498.67			

*For Tabled $F_{1,80}$ = 8.337 at P = .005

$F_{4,80}$ = 2.72 at P = .05

DISCUSSION

RESULTS IN 5 STUDENT SAMPLE

The cbtained $F = 1.81$ was not significant at the $p = .05$ level where a critical value of $F_{9,40} = 2.124$ was necessary to reject the no difference hypothesis. This suggests that the differences between the sample test scores is not significant at the $p = .05$ level; however, as noted in the five subject case, the value of the posttest correct in every sample is above the 50% value used as one of the decision criteria in the evaluation process. All gains are positive and four of the five are near or above the 20% criteria. (The 17.8% gain of Group A is acceptable because 20% is not exact from a sample of nine items; i.e., the cut off is between 11% and 22%.)

The confidence level for each sample mean differences between posttest and pretest means was computer for $p = .20$ and indicated under the "significance column." Note the differences were significant at the stated level in four out of the five samples.

The rho values leave a great deal to be desired except to suggest that four of the five were positive and three of the five are greater than .50. The gain score model does not lend itself to high reliability scores because the amount of gain effects the ranking on the posttest yet the only concern is that gain in fact takes place (Vargas, 1969; Cox, 1966).

RESULTS IN 10 STUDENT SAMPLE

The obtained $F = 1.27$ was not significant at $p = .05$ suggesting the obtained score from the pretest and posttest do not differ enough to indicate a significant change in behavior as a result of seeing the filmstrip. The two-way ANOVA, however, assumes some independence of the pretest and posttest as measures and results in a significant F 9.92, p = .05, with non-significant F values for between groups or interaction.

IMPLICATIONS

The above results indicate that the samples of five students give results which are within the parameters of decision established for the Project. When samples were increased to N = 10, the standard findings were supported for increasing sample size, i.e., scores were within smaller ranges, variance between groups were reduced and gains were more standardized. Why the parameters of all groups tend to be below the population parameters must be attributed to the random variation present in the samples selected. The resulting low posttest scores of the samples cause the gain scores to be depressed more than may be reflected in the population causing all the gain scores to fall below the 20% criteria. All posttests are above the 50% criterion and all gains are positive and at about magnitude which should lead one to realize some stability has been reached.

CONCLUSIONS

Samples of 5 students seem to be large enough to establish estimates of population parameters within the limits of four out of five times. Larger samples do not add appreciably data or the outcome of decisions obtained from the samples of five.

REFERENCES

Computer Based Project Proposal for Fourth Year. Computer Based Project for Evaluation of Media for the Handicapped, City School District, Syracuse, New York, 1972.

Cohen, J. Statistical Power Analysis for the Behavior Sciences, New York: Academic Press, 1969.

Cox, R. C. and Vargas, J. S. A Comparison of Item Selection Techniques. Paper presented at the National Council for Measurement in Education, Chicago, Illinois, 1966.

Gilford, J. P. *Fundamental Statistics in Psychology and Education*, New York: McGraw-Hill, 1965.

Hays, W. L. *Statistics For Psychologists,* New York: Holt, Rinehart, and Winston, 1966.

May, M. L. Y. and Barcikowski. Item Sampling Option: Number of People and Items. Paper presented at the meeting of the American Educational Research Association, New Orleans, February, 1973.

Owen, B. *Handbook of Statistical Tables*. Reading, Mass: Addison Wesley, 1962.

Shoemaker, D. M. A Note on Allocating Items to Subtests on Multiple Matrix Sampling and Approximating Standard Errors of Estimate with the Jackknife. Paper presented at the meeting of the American Educational Research Association, April, 1973.

Siegel, S. *Non-Parametric Statistics For The Behavioral Sciences*. New York: McGraw-Hill, 1956.

Vargas, J. S. Item Selection for Pretest-Posttest Situations. Morgantown: West Virginia University, 1969.

Winer, B. J. *Statistical Principles in Experimental Design*. New York: McGraw-Hill, 1962.

JAB:mrb
873