(4)

CRM 88-123 / December 1988

RESEARCH MEMORANDUM

A COMPARISON OF LINEAR AND EQUIPERCENTILE TEST EQUATING PROCEDURES IN LARGE SAMPLES

D. R. Divgi

A Division of CNA Hudson Institute

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

Work conducted under contract N00014-87-C-0001

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for Public Release; Distribution unlimited
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) CRM 88-123	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Center for Naval Analyses 6b. OFFICE SYMBOL (If applicable) CNA	7a. NAME OF MONITORING ORGANIZATION Commanding General, Marine Corps Combat Development Command
6c. ADDRESS (City, State, and ZIP Code) 4401 Ford Avenue Alexandria, Virginia 22302-0268	7b. ADDRESS (City, State, and ZIP Code) Warfighting Center Quantico, Virginia 22134
8a. NAME OF FUNDING ORGANIZATION Office of Naval Research 8b. OFFICE SYMBOL (If applicable) ONR	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER N00014-87-C-0001
8c. ADDRESS (City, State, and ZIP Code) 800 North Quincy Street Arlington, Virginia 22217	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 65153M; PROJECT NO. C0031; TASK NO.; WORK UNIT ACCESSION NO.
11. TITLE (Include Security Classification) A Comparison of Linear and Equipercentile Test Equating Procedures in Large Samples	
12. PERSONAL AUTHOR(S) D.R. Divig	
13a. TYPE OF REPORT Final 13b. TIME COVERED FROM TO	14. DATE OF REPORT (Year, Month, Day) December 1988 15. PAGE COUNT 12
16. SUPPLEMENTARY NOTATION	

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Aptitude tests, ASVAB (Armed services vocational aptitude battery), Operational test and evaluation, Scoring, Statistical analysis, Statistical processes, Statistical samples, Test scores, Validation
05	08		
12	03		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Scores on new forms of a test are equated to those on an old form. Two common equating procedures are linear and equipercentile. Cross-validation is used to show that, with sample sizes of 6,500 and above, equipercentile equating is preferable to linear for the Armed Forces Vocational Aptitude Battery.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED / UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED	
22a. NAME OF RESPONSIBLE INDIVIDUAL Colonel Preston	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL

CNA

CENTER FOR NAVAL ANALYSES

A Division of Hudson Institute 4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268 • (703) 824-2000

30 December 1988

MEMORANDUM FOR DISTRIBUTION LIST

Subj: Center for Naval Analyses Research Memorandum 88-123

Encl: (1) CNA Research Memorandum 88-123, *A Comparison of Linear and Equipercentile Test Equating Procedures in Large Samples*, by D. R. Divgi, December 1988

1. Enclosure (1) is provided as a matter of possible interest.

2. New forms of the Armed Services Vocational Aptitude Battery undergo initial operational test and evaluation (IOT&E), during which the new forms and the reference form 8a are administered to large samples of military applicants. The data provide equatings of new forms to form 8a. Cross-validation is used to compare linear and equipercentile procedures for test equating. The equipercentile procedure is found to be preferable when samples are as large as those in IOT&E.

Lewis R. Cabe
Director
Manpower and Training Program

Distribution List:
Reverse Page

Subj: Center for Naval Analyses Research Memorandum 88-123

Distribution List

SNDL

A1	ASSTSECNAV MRA
A1	DASN MANPOWER (2 copies)
A2A	CNR
A6	HQMC MPR
	Attn: M
	Attn: MP
	Attn: MR
	Attn: MA (2 copies)
	Attn: MPP-39
A6	HQMC RA
A6	HQMC AVN
A6	CG MCRDAC, Washington
FF38	USNA
	Attn: Nimitz Library
FF42	NAVPGSCOL
FF44	NAVWARCOL (2 copies)
FJA1	COMNAVMILPERSCOM
FJB1	COMNAVCRUITCOM
FKQ6D	NAVPERSRANDCEN
	Attn: Technical Director (Code 01)
	Attn: Director, Testing Systems (Code 63)
	Attn: Technical Library
	Attn: Director, Personnel Systems (Code 62)
	Attn: CAT/ASVAB PMO
	Attn: Manpower Systems (Code 61)
FT1	CNET
V12	CG MCRDAC, Quantico
	Attn: Director, Development Center Plans Division (Code D08) (2 copies)
	Attn: Commanding General
V12	CGMCCDC
	Attn: Training and Education Center

OPNAV

OP-01
OP-11
OP-13
OP-15

OTHER

Joint Service Selection and Classification Working Group (13 copies)
Defense Advisory Committee on Military Personnel Testing (8 copies)

CRM 88-123 / December 1988

A COMPARISON OF LINEAR AND EQUIPERCENTILE TEST EQUATING PROCEDURES IN LARGE SAMPLES

D. R. Divgi

A Division of CNA Hudson Institute

CENTER FOR NAVAL ANALYSES

4401 Ford Avenue • Post Office Box 16268 • Alexandria, Virginia 22302-0268

ABSTRACT

Scores on new forms of a test are equated to those on an old form. Two common equating procedures are linear and equipercentile. Cross-validation is used to show that, with sample sizes of 6,500 and above, equipercentile equating is preferable to linear for the Armed Services Vocational Aptitude Battery.

EXECUTIVE SUMMARY

The Armed Services Vocational Aptitude Battery (ASVAB) is used for selection and classification of enlisted personnel. New forms of the ASVAB are developed about every four years, and equated to the reference form 8a. The ideal outcome is that, during operational use of the ASVAB, the distribution of standard scores is the same for all forms. Equating for operational use is based on data collected during Initial Operational Test and Evaluation (IOT&E); sample sizes exceed 10,000 per form.

Two equating procedures often used by psychometricians are equipercentile and linear. When samples are small, the equipercentile method has large random errors. Linear equating is more stable—that is, it has smaller random error. However, it suffers from bias, i.e., systematic errors at high and/or low scores, if the two forms have score distributions with different shapes. Linear equating was used for forms 11, 12, and 13 and for all subtests except one in forms 15, 16, and 17.

As sample size increases, the superior stability of linear equating becomes less important while its bias remains the same. The question addressed in this paper is whether IOT&E samples are large enough to make equipercentile equating preferable to linear. For equipercentile equating in this study, score frequencies were smoothed by a five-point rolling average and a "dogleg" was used—i.e., the equating curve below the fifth percentile was replaced by a straight line.

DATA

Data used in this study were collected from November 1987 to January 1988 during the IOT&E of ASVAB forms 15, 16, and 17. They were provided to CNA by the Air Force Human Resources Laboratory, after some editing to remove errors such as incorrectly coded form numbers. The sample sizes varied from 13,010 for form 17b to 14,963 for form 15a.

METHODOLOGY

For each form the available sample was split into two random, almost equal parts. One part, which will be called the calibration sample, was used for equating; the other part, called the validation sample, was used to evaluate the results of the equating procedures.

The equipercentile method was applied to the validation samples. The resulting standard scores were used as the criterion. For a specific new form, say 15a, the difference between the criterion standard score and the value from linear equating was squared, and averaged over all applicants in the validation sample for form 15a. The square root of this average is the root mean square difference (RMSD) between the linear equating and the criterion. RMSD for equipercentile equating was computed the same way. For any given form of a subtest, the method with smaller RMSD was considered to have performed better.

RESULTS AND CONCLUSIONS

The RMSD values show that the equipercentile method cross-validated better in a large majority of cases. The equipercentile method is superior in 51 of 60 comparisons. If the two methods work equally well, each has a 0.5 chance of having a lower RMSD. Under this null hypothesis, the chance of one method coming out superior in 51 of 60 cases is less than 0.0001. Thus, the results represent true superiority of the equipercentile method, and are not a chance occurrence.

For ASVAB forms 15, 16, and 17, equipercentile equating is preferable to linear with sample sizes of 6,500 to 7,000, and hence even more so with the larger samples available in IOT&E. This conclusion will remain valid for future editions as well unless much greater effort is made to make new forms parallel to form 8a.

TABLE OF CONTENTS

INTRODUCTION

The Armed Services Vocational Aptitude Battery (ASVAB) is used for selection and classification of enlisted personnel. It contains ten subtests: General Science (GS), Arithmetic Reasoning (AR), Word Knowledge (WK), Paragraph Comprehension (PC), Numerical Operations (NO), Coding Speed (CS), Auto and Shop Information (AS), Mathematics Knowledge (MK), Mechanical Comprehension (MC), and Electronics Information (EI). The Verbal (VE) subtest is defined as the sum of WK and PC. Standard scores rather than raw scores on the subtests are used in all decisions based on the ASVAB. Standard scores are integers from 20 to 80, with mean 50 and standard deviation 10 in the 1980 reference population.

New forms of the ASVAB are developed about every four years, and equated to the reference form 8a. The ideal outcome is that, during operational use of the ASVAB, the distribution of standard scores is the same for all forms. Therefore, two scores on different forms of a subtest are equivalent if they have equal percentile ranks in the population of examinees. This is the definition of equipercentile equating [1].

Only a sample of examinees, rather than the entire population, is available in practice. If the sample is small, the random error of equipercentile equating may be unacceptably large. A popular alternative is linear equating, which is more stable—that is, it has much less random error—because it is based only on means and standard deviations of the two forms. However, to the extent that the score distributions of the forms have different shapes, linear equating suffers from bias, i.e., systematic errors, especially at very high and/or low scores.

The choice between linear and equipercentile methods depends on one's judgment about the relative importance of random and systematic error. If the sample is very large, the bias of linear equating exceeds its superiority in random error, and hence the equipercentile procedure is preferable. The opposite is true when the sample is small. The difference between new and old forms determines the "break-even" sample size at which the bias of linear equating just cancels its superior stability against random error. Equipercentile equating is superior above this sample size, which depends on the differences between old and new forms. Suppose the old and new forms of AS are nearly parallel, whereas those of MC differ substantially. Bias is a more serious concern for MC than for AS; therefore, the break-even sample size is smaller.

In practice, of course, the true differences between forms are unknown, and hence so is the break-even sample size. What one can do is to find out which procedure has worked better in the past with the sample sizes available. A new set of ASVAB forms remains operational for about four years. The equating used during this period is based on data from the Initial Operational Test and Evaluation (IOT&E), which has sample sizes of more than 10,000 per form.

The linear procedure was used for forms 11, 12, and 13 and for all subtests except MC in forms 15, 16, and 17 [2]. The purpose of this paper is to demonstrate that, with the large samples available in IOT&E, equipercentile equating is preferable to linear equating.

EQUATING PROCEDURES

In the equipercentile method, score frequencies were smoothed with a five-point rolling average using the weights −3/35, 12/35, 17/35, 12/35, and −3/35 given by Angoff ([1], p. 516), with the following exceptions. Frequencies of zero and perfect scores were left unchanged; those of scores 1 and $n-1$ were replaced by three-point averages with weights 1/4, 1/2, 1/4 (n being the number of items). In addition, to reduce random error at low scores, a "dogleg" [3] was used: the equating curve at the fifth percentile was connected to the point (−.5,−.5) with a straight line.

The linear equating of this study was the standard procedure using means and standard deviations [1], with converted raw scores constrained to lie between −.5 and $n + .5$.

In both equating procedures, raw score equivalents on form 8a were converted to the standard score scale by linear transformation [4]. The values were not rounded to integers because rounding adds noise to the data. Standard scores below 20 were replaced by 20, and those above 80 by 80.

DATA

Data used in this study were collected from November 1987 to January 1988 during the IOT&E of ASVAB forms 15, 16, and 17. They were provided to CNA by the Air Force Human Resources Laboratory, after some editing to remove errors such as incorrectly coded form numbers. Because of an error in one item, MK form 15b data collected in November were discarded. Apart from this, the sample size was the same for all subtests in a given form. The sample sizes varied from 13,010 for form 17b to 14,963 for form 15a.

METHODOLOGY

Ideally, an equating based on the IOT&E should be evaluated using the subsequent operational data. When such data are not in hand, one can use cross-validation. Six new ASVAB forms are constructed at one time. Thus, during IOT&E of forms 15, 16, and 17, six new forms and form 8a were administered to equivalent samples of applicants to the military services. For each form the available sample was split into two random, almost equal parts. One part, which will be called the calibration sample, was used for equating; the other part, called the validation sample, was used to evaluate the results of the equating procedures.

The basic question is whether, in the validation samples, standard scores on old and new forms have identical distributions. In principle, this can be addressed directly by examining cumulative distributions of standard scores. In practice, however, this leads to serious difficulties because a given raw score is converted into different standard scores for different forms.

A simpler approach is to apply the equipercentile method to the validation samples, and compare the resulting standard scores with those obtained from the calibration samples.

Standard scores obtained from the validation samples were used as the criterion. (To avoid biasing the analysis in favor of equipercentile equating, neither smoothing nor dogleg was used in the criterion equating.) Denote the criterion standard scores by SS_C. Let SS_L and SS_E be standard scores obtained by applying the linear and equipercentile procedures to the calibration samples. For a specific new form, say 15a, the difference $(SS_L - SS_C)$ was squared and averaged over all applicants in the validation sample for form 15a. The square root of this average is the root mean square difference (RMSD) between the linear equating and the criterion. (This statistic is similar in spirit but not in detail to that used by Kolen [5].) RMSD for equipercentile equating was computed the same way. For any given form of a subtest, the method with smaller RMSD was considered to have performed better.

Another summary statistic is the average absolute difference (AAD). It is obtained by computing the mean of the absolute value of the difference. Again, a smaller AAD represents better performance.

RESULTS AND CONCLUSIONS

Table 1 presents the RMSD values for all forms of all subtests. They show that the equipercentile method cross-validated better in a large majority of cases. If we exclude MC, for which linear equating has already been found to be inadequate [2], the equipercentile method is superior in 51 of 60 comparisons. If the two methods work equally well, each has a 0.5 chance of having a lower RMSD. Under this null hypothesis, the chance of one method coming out superior in 51 of 60 cases is less than 0.0001.

Table 2 presents the AAD values. Again the superiority of the equipercentile method is evident, with AAD for the equipercentile being smaller in 52 of the 60 cases excluding MC.

Note that the equatings were carried out with half the IOT&E sample. Thus, with sample sizes around 6,500 to 7,000, the equipercentile method turns out to be preferable to linear equating. The superiority of the former will be even more striking with the full IOT&E samples because, as sample size increases, the superior stability of linear equating becomes less important while its bias remains the same. How does the bias of linear equating depend on raw scores? Results of simulations show that bias is minimal near the mean score, and large at high and low scores [6].

The relative merits of the two methods also depend on the degree to which old and new forms differ. When new forms of the ASVAB are developed, efforts are made to make them parallel to the reference form by careful selection of items from overlength versions of the new forms. Some differences remain, due to the limited sizes of the overlength forms and of the recruit samples. Unless these are increased substantially, future ASVAB forms will differ from form 8a to roughly the same extent as forms 15, 16, and 17; hence, the conclusion of this paper will remain applicable.

Table 1. Root mean square change in standard score from equating sample to validation sample

Equating procedure	Form					
	15a	15b	16a	16b	17a	17b
General Science						
Linear	.280	.445	.497	.470	.491	.471
Equipercentile	.171	.317	.243	.247	.273	.288
Arithmetic Reasoning						
Linear	.222	.403	.517	.302	.312	.479
Equipercentile	.297	.286	.406	.264	.226	.322
Word Knowledge						
Linear	.475	.436	.330	.328	.371	.421
Equipercentile	.285	.137	.286	.185	.218	.279
Paragraph Comprehension						
Linear	.517	.356	.604	.223	.389	.371
Equipercentile	.259	.138	.156	.147	.251	.307
Numerical Operations						
Linear	.242	.206	.432	.226	.172	.442
Equipercentile	.122	.321	.233	.244	.132	.409
Coding Speed						
Linear	.364	.252	.316	.450	.358	.178
Equipercentile	.301	.196	.305	.480	.365	.230
Auto and Shop Information						
Linear	.166	.436	.567	.379	.309	.364
Equipercentile	.134	.450	.425	.337	.251	.287
Mathematics Knowledge						
Linear	.304	.344	.199	.202	.397	.369
Equipercentile	.189	.255	.177	.200	.284	.216
Mechanical Comprehension						
Linear	.640	.671	.780	.812	.723	.741
Equipercentile	.251	.323	.274	.336	.223	.280
Electronics Information						
Linear	.622	.536	.176	.125	.315	.466
Equipercentile	.218	.291	.220	.214	.255	.271
Verbal						
Linear	.502	.313	.320	.361	.474	.387
Equipercentile	.238	.147	.244	.233	.405	.262

Table 2. Average absolute change in standard score from equating sample to validation sample

Equating procedure	Form					
	15a	15b	16a	16b	17a	17b
General Science						
Linear	.183	.370	.404	.381	.444	.395
Equipercentile	.155	.264	.202	.219	.238	.230
Arithmetic Reasoning						
Linear	.172	.350	.385	.259	.270	.407
Equipercentile	.275	.228	.370	.214	.165	.287
Word Knowledge						
Linear	.351	.291	.217	.245	.245	.262
Equipercentile	.214	.088	.214	.126	.169	.212
Paragraph Comprehension						
Linear	.361	.300	.547	.194	.295	.231
Equipercentile	.201	.118	.095	.132	.093	.268
Numerical Operations						
Linear	.148	.134	.376	.151	.143	.367
Equipercentile	.074	.179	.173	.129	.084	.331
Coding Speed						
Linear	.308	.208	.245	.394	.217	.141
Equipercentile	.240	.142	.218	.367	.230	.170
Auto and Shop Information						
Linear	.128	.394	.413	.304	.247	.271
Equipercentile	.105	.406	.364	.204	.223	.224
Mathematics Knowledge						
Linear	.265	.279	.165	.164	.344	.322
Equipercentile	.164	.216	.147	.144	.260	.190
Mechanical Comprehension						
Linear	.496	.569	.668	.662	.480	.567
Equipercentile	.189	.289	.228	.261	.125	.240
Electronics Information						
Linear	.507	.443	.139	.087	.251	.397
Equipercentile	.180	.250	.192	.177	.219	.240
Verbal						
Linear	.444	.227	.254	.288	.355	.301
Equipercentile	.184	.112	.212	.187	.265	.223

REFERENCES

[1] W. H. Angoff. "Scales, Norms, and Equivalent Scores," in *Educational Measurement*, edited by Robert L. Thorndike. 2nd ed. Washington, D.C.: American Council on Education, 1971

[2] Linda Curran. *IOT&E of ASVAB Forms 15/16/17*, Briefing presented to the Joint Service Selection and Classification Working Group, Jun 1988

[3] Henry I. Braun and Paul W. Holland. "Observed-Score Test Equating: A Mathematical Analysis of Some ETS Equating Procedures." In *Test Equating*, edited by Paul W. Holland and Donald B. Rubin. New York: Academic Press, 1982, 9-49

[4] CNA Report 116, *The ASVAB Score Scales: 1980 and World War II*, by Milton H. Maier and William H. Sims, Jul 1986 (94011600)[1]

[5] Michael J. Kolen. "Comparison of Traditional and Item Response Theory Methods for Equating Tests," *Journal of Educational Measurement* (Spring 1981): 1-12

[6] CNA Research Contribution 571, *A Stable Curvilinear Alternative to Linear Equating*, by D. R. Divgi, Oct 1987 (02057100)

1. The number in parentheses is a CNA internal control number.