DOCUMENT RESUME

ED 128 414 TM 005 598

AUTHOR Sands, William A.
TITLE Alternative Item Response Weighting Procedures: Development and Evaluation.
PUB DATE [Sep 75]
NOTE 16p.; Paper presented at the Annual Conference of the Military Testing Association (17th, Fort Benjamin Harrison, Indiana, September 15-19, 1975); Also included in TM 005 585

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Admission (School); College Majors; College Students; *Comparative Analysis; Educational Guidance; *Interest Tests; *Military Personnel; Occupational Guidance; Predictive Ability (Testing); *Scoring Formulas; Statistical Analysis; *Test Interpretation; Test Validity; Vocational Interests; *Weighted Scores
IDENTIFIERS Naval Academy; *Strong Vocational Interest Blank

ABSTRACT

In order to develop tools for use in the selection and vocational-educational guidance of U.S. Naval Academy midshipmen, three empirically-based scales, designed using the Strong Vocational Interest Blank (SVIB), were developed to predict three criteria: (1) disenrollment for academic reasons, (2) disenrollment for motivational reasons, and (3) military aptitude. The Naval Academy classes of 1971, 1972, and 1973 took the SVIB, and an empirical criterion keying approach was used to select those items having the 75 best responses for each of four different academic major groupings. Twenty alternative item response weighting methods were evaluated. For each of the four problems, a number of different response weighting methods had essentially the same effectiveness. A parsimonious conclusion would suggest the continued use of the common procedure of assigning positive or negative unit weights to the responses. However, scale test-retest reliability and scoring costs are two pertinent factors which should be included in an overall evaluation of alternative item response weighting procedures for a particular application. (BW)

ALTERNATIVE ITEM RESPONSE WEIGHTING PROCEDURES: DEVELOPMENT AND EVALUATION[1]

by

William A. Sands

Navy Personnel Research and Development Center
San Diego, California 92152

BACKGROUND AND PURPOSE

The investigation described herein was accomplished under a program at the Navy Personnel Research and Development Center, San Diego, CA. The major purpose of this research program is the development of tools for use in selection and vocational-educational guidance for the U. S. Naval Academy midshipmen.

A rather substantial amount of research has been done on interests and their relationship to various criteria (Campbell, 1966). Using the Strong Vocational Interest Blank (SVIB), Abrahams and Neumann (1973) developed empirically-based SVIB scales designed to predict three criteria for the U. S. Naval Academy: (1) disenrollment for academic reasons; (2) disenrollment for motivational reasons; and (3) military aptitude. All three new scales evidenced significant relationships with their respective criteria in cross-validation samples.

The Naval Academy "Majors Program" was initiated in 1969 and subsequently revised. At present, academic majors are organized under three broad areas (I) Engineering-Weapons; (II) Mathematics-Science; and (III) Humanities.

Recently, there has been a marked emphasis on the importance of the technical majors included in Groups I and II. The 1974 edition of the Majors Program, published by the Academy, outlines current policy:

> The Naval Academy policy on the selection of majors is clear. Each midshipman selects a major which will meet the needs of the Navy and at the same time be

---

[1] The opinions expressed are solely those of the author and do not necessarily reflect official Department of the Navy policy. Reproduction in whole or in part is permitted for any purpose of the United States Government.

> interesting to him. The needs of the Navy take first priority and it has been determined that eighty percent of the Class shall take a technical major, i.e, Group I or II, and twenty percent may choose Group III. Hopefully, the selection of majors by the Class of 1978 will meet the 80/20 quota. If the desired distribution is not obtained by an open, free selection process, steps will be taken to adjust the distribution to meet the Navy goals. (U.S. Naval Academy, p. 9)

Pursuant to this emphasis on technical majors, Neumann and Abrahams (1974) developed a SVIB scale (E-S) designed to identify Naval Academy applicants with engineering and science interests. The class of 1973 was split into a key-development sample and a cross-validation sample. The criterion employed was dichotomous: Engineering-Science major versus "other" major. A biserial validity of .57 was obtained in the cross-validation sample. Application of the new SVIB scale to the class of 1976 yielded a biserial validity of .62. The results of this investigation were also reported in a paper presented at an Air Force symposium (Abrahams & Neumann, 1974).

The Engineering-Science scale (E-S) was developed for use in the selection of students from the pool of Academy applicants. The current operational selection composite involves a number of different predictors. The relationship between the E-S scale and the current predictors was examined and the validity of alternative composites was evaluated against various criteria; e.g., cumulative grade point average, major choice, etc. To facilitate utilization of the research findings, results have been forwarded by letter to the Dean of Admissions of the Naval Academy (Neumann, 1975). Both the Disenrollment scale and the Engineering-Science scales will be employed in computing the candidate multiple for applicants for the class of 1980 (McKee, 1975).

PROCEDURE

Instrument

The 1966 edition of the Strong Vocational Interest Blank (SVIB) for men contains 399 items dealing with occupational activities, school subjects, etc. A person taking the SVIB is asked to endorse one of three response alternatives for most of the items: "Like," "Indifferent," or "Dislike."

Samples

The U. S. Naval Academy classes of 1971, 1972, and 1973 took the SVIB during their respective plebe summers. The sample data were edited to remove persons who failed to graduate and persons who graduated with a major in the management science area. The number of persons in each specific major for each class is shown in Table 1.

Criteria

Graduates in the specific majors were aggregated into broad areas, as shown in the bottom of Table 1. Four separate problems were addressed: (1) differentiation of Engineering-Weapons majors (Group I) from all other (Groups II + III); (2) differentiation of Mathematics-Science majors (Group II) from all others (Groups I + III); (3) differentiation of Humanities majors (Group III) from all others (Groups I + II); and (4) differentiation of Engineering-Weapons majors (Group I) from Mathematics-Science majors (Group II). Previous research (Sands & McCullah, 1974) has indicated that separating Group I persons from Group II persons on the basis of their SVIB responses is more difficult than differentiating Group III majors from all others.

Item Selection

An "empirical criterion keying" approach was used to select those SVIB items having the 75 best responses for each of the four problems addressed. The proportion of high criterion group members who endorsed each of the response alternatives for each of the items was computed. The same proportion was computed for the low criterion group. Then the absolute difference between these two endorsement rates was computed. The items containing the 75 responses exhibiting the greatest absolute differences between endorsement rates were selected for subsequent weighting.

Response Weighting

Twenty alternative item response weighting methods were evaluated. In each method, weights were assigned so that high scores would be associated with the high criterion group, while the low criterion group would tend to receive lower scores.

Method #1. Unit weights were assigned to the 75 responses with the greatest absolute difference in endorsement rates. For those responses endorsed by a greater proportion of high criterion group members than by low criterion group members, a positive unit weight was assigned. Conversely, a negative unit weight was given to responses endorsed by a greater proportion of low criterion group members. Responses which were not among the best 75 received weights of zero.

Method #2. Those items having one or more responses receiving a unit weight under the first method were dimensionalized, as suggested by Campbell (1971). Each such item is considered as a continuum ranging

TABLE 1

Number of Naval Academy Graduates by Individual Major for the Classes of 1971, 1972 and 1973

Individual Major	Number of Graduates 1971	1972	1973	Total
Group I: Engineering-Weapons				
Aerospace Engineering	76	51	38	165
Electrical Engineering	22	24	22	68
General Engineering	0	0	16	16
Marine Engineering	10	13	4	27
Mechanical Engineering	66	41	37	144
Naval Architecture	15	6	10	31
Ocean Engineering	13	18	32	63
Systems Engineering	23	16	12	51
Group II: Mathematics-Science				
Applied Science	12	19	18	49
Chemistry	12	14	35	61
Mathematics	85	58	66	209
Oceanography	95	122	103	320
Operations Analysis	31	51	63	145
Physical Science	0	0	1	1
Physics	29	43	28	100
Group III: Humanities				
American Political Systems	25	49	48	122
International Security Affairs	53	54	56	163
European Studies - French	5	4	7	16
European Studies - German	15	7	3	25
European Studies - Italian	2	1	0	3
Latin American Studies - Spanish	15	12	7	34
Latin American Studies - Portuguese	0	0	2	2
Far Eastern Studies - Chinese	0	3	1	4
Soviet Studies - Russian	2	2	3	7
Economics	23	12	11	46
English	12	6	10	28
History	19	20	30	69

Summary	1971 N	1971 %	1972 N	1972 %	1973 N	1973 %	Total N	Total %
Group I	225	34	169	26	171	26	565	29
Group II	264	40	307	48	314	47	885	45
Group III	171	26	170	26	178	27	519	26

from "Like" at one extreme to "Dislike" at the opposite end. If one end of the continuum received a unit weight under the first weighting procedure, that response receives the same weight under this method and the opposite end of the response continuum is assigned a unit weight affixed with the opposite sign. If "Indifferent" was the only response for an item which received a unit weight under the first weighting procedure, the item obviously does not have the assumed underlying continuum and, therefore, all responses for the item are assigned a weight of zero under this second weighting method.

Method #3. Multiple weights were chosen after examination of the distribution of absolute differences between endorsement rates for the best 75 responses. These weights (0, ± 1, ± 2, ± 3, ± 4 and ± 5) were assigned to responses according to the degree to which the two criterion groups differed. Again, positive weights were attached to responses endorsed by a greater proportion of the high criterion group than the low criterion group. Negative multiple weights were assigned to responses endorsed by a greater proportion of low than high criterion group members. The positive and negative multiple weights were assigned only to those responses receiving a unit weight under the first method. Those responses receiving a zero weight under the first method also receive a zero weight under this third method.

Method #4. The fourth weighting method is a dimensionalized version of the third method. The assumption and procedure employed in dimensionalizing was explained under the second weighting method discussed above.

Method #5. Examination of the results of other investigations suggested another set of multiple weights. These weights (0, ± 3, ± 4, ± 5, and ± 6) were assigned using essentially the same procedure as explained above for the third method. Again, weights were assigned only to those responses which received a unit weight under the first method. A weight of zero was assigned to those responses having zero weights under the first method.

Method #6. The sixth response weighting method is simply a dimensionalized version of the fifth method.

Method #7. This response weighting method assigns weights based upon the endorsement rate for the high criterion group, regardless of the low criterion group endorsement rate.

Method #8. The eighth method uses weights based upon the differences between the endorsement rates for the high and low criterion groups.

Method #9. Like the previous response weighting method, this ninth method employs the difference between the endorsement rates of the high and low criterion groups. Under the present method, this difference is squared and affixed with the sign of the nonsquared difference.

Method #10. The tenth response weighting method utilized Bayes' Theorem to estimate the probability that a person belongs to the high criterion group, given that he endorsed a particular response alternative. The desired posterior probability is a function of the conditional probabilities of endorsing the response, given membership in each of the criterion groups and the prior probabilities of belonging to each of the criterion groups.

Method #11. The posterior probability of being a member of the high criterion group, computed under the tenth method, may be considered as a proportion. Specifically, of all the persons who endorsed a particular alternative of an item, a certain proportion belong to the high criterion group. Each proportion has a standard error which is a function of the proportion itself and the sample size. This eleventh method involves inversely weighting the proportion by the standard error of the proportion.

The standard error of a proportion is influenced by the proportion itself and the sample size for the response. Specifically, the inverse of a standard error of a proportion is smallest when the proportion is 0.5 and becomes progressively larger as the proportion approaches zero or unity. This means that, for a fixed sample size, extreme proportions are weighted by a higher factor. The product of this factor (the inverse of the standard error of a proportion) and the proportion itself yields a very large weight for high proportions in comparison to the weight for low proportions, for a fixed sample size.

On the other hand, for a fixed proportion, a larger weight is assigned to those responses endorsed by a large number of persons than is given to responses made by a small number of persons. This characteristic of the eleventh weighting method reflects the belief that a proportion computed in a large sample should be more stable than would be the case for a small sample and, therefore, should receive a higher weight.

Method #12. The proportion considered in the previous two methods was weighted by a factor consisting of the ordinate on the unit normal curve divided by the standard error of the proportion. The weight assigned to a response under this twelfth method is influenced by the proportion and the sample size. As was true for the eleventh method, responses endorsed by a large number of persons receive more weight than responses based upon a small sample, when the proportion is held constant. The eleventh method modifies the original proportion by assigning larger weights to extreme proportions for a fixed sample size.

This twelfth method reverses this strategy. The original proportion is modified by assigning larger weights to proportions near 0.5 and progessively smaller weights to extreme proportions near zero or unity. The original proportion is weighted most heavily when the response distribution is equally divided by the two criterion groups.

Method #13. As mentioned above, the posterior probability of being a member of the high criterion group, given the endorsement of a particular response, is a proportion. This thirteenth method divides this proportion by its complement and performs a natural logarithmic transformation on the result.

Method #14. A two-by-two contingency table was constructed for each response alternative for each item evaluated. The rows of the tables represent the two criterion groups (high and low) while the columns represent a dichotomized response (absence versus presence).

A phi coefficient was computed for each response of each item evaluated. This coefficient is used for the fourteenth response weighting method.

Method #15. The coefficient of determination for a validity coefficient is the square of the correlation. This coefficient of determination represents the proportion of the criterion variance which is explained by the predictor variable. The phi coefficient (the validity for the dichotomized response, two criterion group problem) was squared and affixed with the sign of the nonsquared difference.

Method #16. The sixteenth method uses the phi coefficient computed for the fourteenth method and inversely weights the coefficient by the standard error of the phi coefficient.

Method #17. The magnitude of any phi coefficient is constrained by the marginal proportions of the two-by-two contingency table. In an attempt to eliminate this constraining influence on the response weights, the seventeenth method uses the ratio of the obtained phi coefficient to the maximum phi coefficient possible under the given conditions.

Method #18. The eighteenth response weighting method employs Fisher's Z coefficient, a transformation of the correlation coefficient. For two predictor categories (presence versus absence of the response) and two criterion categories (high versus low), the phi coefficient is the appropriate correlation.

Method #19. The nineteenth method uses the square of the Fisher's Z coefficient computed for the previous method. This squared coefficient is affixed with the sign of the original Z coefficient.

Method #20. The last method weights the Fisher's Z coefficient, computed for the eighteenth method, by the inverse of the standard error of the coefficient.

Scale Evaluation

The SVIB item responses for each member of the Classes of 1971, 1972, and 1973 were scored for each of the eighty scales for all midshipmen. Means and standard deviations were computed separtely for the high and low criterion groups for all scales. Point-biserial validity coefficients were computed separately for each year group, for each of the eighty scales. Finally, for each separate scale, two validities obtained in the Class of 1971 and the Class of 1972 were averaged. Each validity coefficient was transformed into a Fisher's Z coefficient, weighted by the appropriate degrees of freedom, averaged, and then converted back to a correlation. This weighted average cross-validity was used to assess the effectiveness of each of the twenty alternative item response weighting procedures. The point-biserial validities for the Class of 1973 were not used to evaluate the twenty methods, as the scales were constructed on this group.

RESULTS

Engineering-Weapons Majors Versus Other Majors

Table 2 shows the point-biserial validity coefficients for each of the twenty alternative scales for each of the three classes. These scales are designed to differentiate persons with a major in the Engineering-Weapons area from persons selecting a major in either of the other two broad academic areas.

The last column in Table 2 presents a weighted average cross-validity for each of the twenty scales. This validity is based upon the Classes of 1971 and 1972. Scales E16 and E20 demonstrated the highest average validity. A number of other scales showed closely similar validities. Specifically, eleven of the twenty scales had weighted average cross-validities within .002 of the highest average validity.

Mathematics-Science Majors Versus Other Majors

Results for the scales designed to differentiate between persons majoring in the Mathematics-Science area and persons majoring in another area are shown in Table 3. The highest weighted average validity of the twenty scales was obtained by the M04 scale. The M17 and M19 scales demonstrated weighted average cross-validities within .002 of the highest one.

TABLE 2

Effectiveness of Twenty Engineering-Science Scales

Scale	Point-Biserial Validity Coefficients: Key Development Sample-1973	Cross-Validation Samples 1971	Cross-Validation Samples 1972	Weighted Average Cross-Validity
E01	.419	.379	.299	.340
E02	.397	.386	.307	.348
E03	.416	.379	.304	.342
E04	.396	.384	.310	.348
E05	.419	.379	.301	.341
E06	.399	.385	.308	.347
F07	.275	.265	.245	.255
E08	.415	.380	.309	.345
E09	.410	.378	.309	.344
E10	.419	.382	.310	.347
E11	.304	.299	.265	.282
E12	.352	.343	.296	.320
E13	.364	.345	.298	.322
E14	.416	.384	.311	.348
E15	.413	.383	.309	.347
E16	.417	.384	.312	.349
E17	.408	.381	.312	.347
E18	.416	.384	.311	.348
E19	.413	.384	.311	.348
E20	.417	.384	.312	.349

TABLE 3

Effectiveness of Twenty Mathematics-Science Scales

Scale	Point-Biserial Validity Coefficients: Key Development Sample-1973	Cross-Validation Samples 1971	Cross-Validation Samples 1972	Weighted Average Cross-Validity
M01	.399	.180	.284	.232
M02	.354	.177	.290	.234
M03	.393	.183	.293	.238
M04	.351	.186	.298	.242
M05	.399	.180	.284	.232
M06	.354	.177	.290	.234
M07	.247	.134	.212	.173
M08	.396	.173	.294	.234
M09	.389	.174	.299	.237
M10	.392	.167	.261	.214
M11	.293	.154	.229	.191
M12	.296	.152	.230	.191
M13	.286	.158	.257	.207
M14	.392	.175	.299	.237
M15	.384	.175	.302	.239
M16	.395	.175	.298	.237
M17	.370	.176	.305	.241
M18	.392	.175	.299	.237
M19	.384	.177	.302	.240
M20	.395	.175	.298	.237

Humanities Majors Versus Other Majors

Table 4 presents the point-biserial validities for each class on each of the twenty alternative scales designed to differentiate persons majoring in the area of Humanities from persons majoring in the other two broad academic areas. The highest weighted cross-validity was obtained by the H15 and H19 scales. Three other scales, H17, H18 and H20, demonstrated weighted average cross-validities within .002 of the highest one.

Engineering-Science Majors Versus Mathematics-Science Majors

The last of the four problems addressed in this study was to differentiate persons majoring in the Engineering-Science area from persons majoring in the Mathematics-Science area. The point-biserial validity coefficients for the twenty alternative item response weighting strategies for each of the three classes are presented in Table 5. The T17 scale showed the highest average cross-validity. The average validities obtained for the T04 and T06 scales were within .002 of the highest scale.

DISCUSSION AND CONCLUSIONS

Most Effective Item Response Weighting Methods

The most striking characteristic of the twenty alternative item response weighting methods is their general similarity in terms of effectiveness. For each of the four problems addressed in this study, a number of different response weighting methods have essentially the same ability to differentiate between the high and low criterion groups. A parsimonious conclusion would suggest the continued use of the common procedure of assigning positive or negative unit weights to the responses; i.e., the first method.

Least Effective Item Response Weighting Methods

Unlike the situation for the most effective method where a number of techniques are essentially equivalent, there are two methods which were consistently the least effective. Method #7 ranks as the worst method of all twenty methods. This finding holds true across all four problems. This method employs weights based upon the endorsement rate of the high criterion group, and ignores the low criterion group endorsement rate.

Method #11 was the second least effective method in each of the four problems. This method weighted the posterior probability of high criterion group membership by the inverse of the standard error.

TABLE 4

Effectiveness of Twenty Humanities Scales

Scale	Point-Biserial Validity Coefficients			
	Key Development Sample-1973	Cross-Validation Samples 1971	Cross-Validation Samples 1972	Weighted Average Cross-Validity
H01	.582	.535	.549	.542
H02	.575	.528	.543	.535
H03	.585	.534	.556	.545
H04	.581	.528	.549	.538
H05	.586	.534	.556	.545
H06	.581	.530	.550	.540
H07	.258	.271	.149	.211
H08	.581	.535	.554	.544
H09	.585	.532	.557	.544
H10	.591	.531	.565	.548
H11	.422	.391	.306	.350
H12	.505	.464	.419	.442
H13	.584	.52[illegible]	.546	.534
H14	.587	.537	.562	.549
H15	.590	.532	.572	.552
H16	.587	.536	.562	.549
H17	.585	.528	.572	.550
H18	.586	.537	.563	.550
H19	.590	.531	.572	.552
H20	.587	.537	.563	.550

TABLE 5

Effectiveness of Twenty Technical Scales

Scale	Point-Biserial Validity Coefficients: Key Development Sample-1973	Cross-Validation Samples: 1971	Cross-Validation Samples: 1972	Weighted Average Cross-Validity
T01	.467	.214	.213	.214
T02	.391	.253	.207	.230
T03	.473	.204	.214	.209
T04	.403	.248	.213	.231
T05	.467	.214	.214	.214
T06	.391	.254	.208	.231
T07	.212	.159	.100	.130
T08	.476	.215	.203	.209
T09	.476	.219	.199	.209
T10	.481	.223	.211	.217
T11	.275	.183	.141	.162
T12	.329	.211	.157	.184
T13	.464	.213	.205	.212
T14	.483	.221	.208	.215
T15	.488	.224	.210	.217
T16	.484	.222	.204	.213
T17	.482	.244	.222	.233
T18	.483	.222	.209	.216
T19	.488	.224	.210	.217
T20	.484	.222	.204	.213

Generalization of the Findings

This study was comprehensive in that twenty item response weighting procedures were examined in four separate problems, using three separate samples. However, the extent to which the findings can be generalized is limited by two considerations. The first limitation is that the original "best" 75 responses were selected on the basis of the greatest absolute differences in endorsement rates between the high and low criterion groups. Other procedures for selecting items to be keyed could produce results different from those reported herein.

A more important limitation is that the number of responses originally keyed was not systematically varied. For each of the four problems, those items containing the best 75 responses were chosen for keying. It is expected that one or a few of the more sophisticated differential weighting methods might evidence a distinct advantage over the simple unit weighting method if the number of keyed responses was decreased. Conversely, as key length increases, unit weighting may evidence marked advantages over the more mathematically complex methods.

Finally, the "best" method should not be determined solely on the basis of a validity coefficient. Scale test-retest reliability and scoring costs are two pertinent factors which should be included in an overall evaluation of alternative item response weighting procedures for a particular application.

REFERENCES

1. Abrahams, N. M. and Neumann, I. The validation of the Strong Vocational Interest Blank for predicting Naval Academy disenrollment and military aptitude. San Diego, U. S. Naval Personnel and Training Research Laboratory, February 1973. (Technical Bulletin STB 73-3).

2. Abrahams, N. M. and Neumann, I. Identification of Naval Academy applicants with engineering and science interests. Proceedings: Psychology in the Air Force Symposium. Colorado: U. S. Air Force Academy, April 1974.

3. Campbell, D. P. Manual for the Strong Vocational Interest Blank. Stanford: Stanford University Press, 1966.

4. Campbell, D. P. Handbook for the Strong Vocational Interest Blank. Stanford: Stanford University Press, 1971.

5. McKee, K. R. Personnel Letter to Commanding Officer, Navy Personnel Research and Development Center from Superintendent, U. S. Naval Academy, 25 August 1975.

6. Neumann, I. and Abrahams, N. M. Identification of Naval Academy applicants with engineering and science interests. San Diego: Navy Personnel Research and Development Center, October 1974 (Technical Report 75-7).

7. Neumann, I. Preliminary results of candidate multiple reweighting study. San Diego: Navy Personnel Research and Development Center, April 1975. (Letter to Dean of Admissions, U.S. Naval Academy).

8. Sands, W. A. and McCullah, R. D. Development and evaluation of Strong Vocational Interest Blank scales for academic major areas at the U. S. Naval Academy. Oklahoma City, Oklahoma: Proceedings of the 16th Annual Conference of the Military Testing Association.

9. U. S. Naval Academy. The Majors Program. Annapolis: Academic Dean Notice 1531 of 1 July 1974.