ED 318 791 TM 014 906

AUTHOR Seong, Tae-Je

TITLE Reconsideration of the ANOVA Method of Detecting Item Bias.

PUB DATE Apr 90

NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Ability Identification; *Analysis of Variance; Black Students; *College Students; Higher Education; *Item Bias; Racial Bias; Sample Size; Statistical Bias; *Test Items; *Validity; White Students

IDENTIFIERS Split Plot Designs

ABSTRACT

The validity of the analysis of variance (ANOVA) of a split plot factorial design was investigated using a complex interaction contrast with matched two-group ability data for detecting biased items. The definition of a biased item by this method is: an item is biased if there is an item-by-group interaction when there is no group difference in achievement levels. Data were drawn from a study by M. J. Subkoviak et al. (1984), consisting of the responses of 1,022 white and 1,008 black college students to 50 multiple-choice vocabulary test items. In this data set, bias was deliberately manipulated by including black vocabulary items. Sixty matched pairs of subjects were selected to determine the ANOVA method with a small data group. The ANOVA method with the matched-pair data had a large interaction for an item if that item was biased. The ANOVA method with small data set is still valid for detecting biased items after matching the abilities of group members. Four tables are included and two tables are appended. (SLD)

Reconsideration of the ANOVA method for Detecting Item Bias

Seong, Tae-Je

Ewha Womans University

Paper presented at the annual meeting of the
American Educational Research Association

Boston, April, 1990

Abstract

The purpose of this study was to investigate the validity of the analysis of variance of a split plot factorial design using a complex interaction contrast with the matched-pair data of two group abilities. The definition of a biased item via this method is: an item is defined as biased if there is an item-by-group interaction when there is no group difference in group achievement levels.

This study showed that the ANOVA method with the matched-pair data had a large interaction for an item if an item was biased. The ANOVA method with matched-pair data is capable of detecting a biased item when there are small number of subjects in both groups.

Reconsideration of the ANOVA method for Detecting Item Bias

Background and Purpose

Analysis of variance (ANOVA) is one of the old methods for detecting biased items. According to this method, a biased item is defined as an item-by-group interaction (Clearly & Hilton,1968). The ANOVA method has been criticized by many researchers (Hunter,1975; Lord,1977; Jensen,1980; Camilli & Shepard; 1987). Hunter (1975) discussed that an item of different difficulty shows an item-by-group interaction in any situation in which the two groups differ in achievement level even though there is absolutely no bias in the item toward one group or the other. Camilli and Shepard (1987) insisted that "when there is a true difference in group achievement level, the ANOVA interaction is incapable of detecting bias that adds or subtracts from this true difference". The common criticism for the ANOVA method is that this method tends to misclassify highly discriminating items as biased when two groups have real big different abilities.

Therefore, the Rudner's area measure (1977) using three-parameter item characteristic curve, the Mentel-Haenszel method which was proposed by Holland and Thayer (1988), and the Camilli's chi-square method (1979) have been recommended widely to detect biased items. The Linn and Harnisch's pseudo-IRT(Z) method (1981) and the Camilli's method (1979) are recommendable when a minority group is small (Seong & Subkoviak, 1987; Shepard, Camilli, & William, 1985). But these methods still require large sample size to estimate item and ability parameters and to avoid the violation of the assumption of chi-square statistics. A current study (Seong & Subkoviak, 1987) has pointed out the theoretical limitation of the Linn and Harnisch's pseudo-IRT(Z) method that estimates of item parameters are influenced by combining a minority group with a majority group. The Camilli's chi-square method is sensitive when cutoff scores for the intervals are changed. There is still a concern to find a valid easy item bias detection procedure when there is a small number of subjects regardless of groups.

Regardless of the different item bias detection methods, the widely accepted definition is: an item is biased if individuals with equal ability, but from different groups, have unequal probability of answering the item correctly. This definition should be applied to the interaction concept of the ANOVA method. The serious flaw of the ANOVA method is that a highly discriminating item and an very easy or difficult item are defined as biased when there is real large difference in group abilities. Matching group abilities for the ANOVA method may solve the problem. According to the widely accepted definition of a biased item and the interaction concept of the ANOVA method, an item is considered as biased if there is an item-by-group interaction when there is no group difference in group achievement levels. In other word, an item is defined as a biased item if there is large difference of item difficulties when one group ability is matched to another group ability.

This study investigated the validity of the ANOVA method using a interaction contrast of a split-plot factorial design with matched two group ability data for detecting biased items. There are two reasons to improve the ANOVA method. One reason is that this method does not require a large sample size for both groups. Another reason is that it does not require a sophisticated formula and subjective idea to define biased items.

Methodology

Analysis of Variance

Design A design of analysis of variance for detecting biased items is a split-plot factorial design in which two or more groups completely cross the items as the treatments. The split-plot factorial design is a repeated measure. This model is accomplished by forming i=1,2,.....n blocks of homogeneous groups where the blocks with students represent the levels of the extraneous variable. It is a mixed effect model because item is fixed, group is fixed, and student is a random blocking variable. Student is randomly selected from the population of each group.

Method If an item is not biased, there is no interaction between the group ability and the item difficulty. The interaction for each item means a difference between a difference of the group abilities and a difference of the item difficulties. Therefore, the complex interaction contrast is used to measure a degree of bia: for each item. The complex interaction contrast between the group ability and the item difficulty for each item is below;

$$\hat{\psi}_k = \Big[\frac{1}{n_j(q-1)}(Y_{.j1} + \dots\dots + Y_{.jk-1} + Y_{.jk+1} + \dots\dots + Y_{.jq}) - \frac{1}{n_{j'}(q-1)}(Y_{.j'1} + \dots\dots + Y_{.j'k-1} + Y_{.j'k+1} + \dots\dots + Y_{.j'q}\Big] - \Big[\frac{1}{n}Y_{.jk} - \frac{1}{n}Y_{.j'k}\Big] \qquad (1)$$

where

$Y_{.jk}$ = sum of scores (1,0) on item k in group j;
n_j = total number of examinees in group j; and
q = total number of items.

Procedures to measure a degree of bias for each item through the ANOVA method with the complex interaction contrast between the group ability and the item difficulty are following;

1) get the second error term in the ANOVA table of the split-plot factorial design;
2) compute sum of scores (1,0) on each item within each group;
3) compute an item difficulty in each group;
4) sum total item score for all items except one item and divide the sum of the total score by number of examinees and one less number of items in order to compute group ability in each group;
5) compute the complex interaction contrast by Equation (1);
6) compute a standard error of the interaction contrast by Equation (2);

$$SE\ \hat{\psi}_k = \sqrt{MSE \sum_{k=1}^{q} \frac{c_k^2}{n_j}} \qquad (2)$$

7) compute a bias index for each item by Equation (3);

$$T_k = \frac{\hat{\gamma}_k}{SE\ \hat{\gamma}_k} \qquad (3)$$

T is an index of a degree of an interaction between the group ability and the item difficulty. The T index is zero when an item is not biased; while a large T value (positive or negative) suggests the presence of bias. In this study, a positive sign indicates that an item favors the minority group, because their actual performance is better than their group ability. Unsigned measure is the absolute value of the signed measure.

Data Source

Data for this study were from the study by Subkoviak, Mack, Ironson, and Craig (1984). The purpose of using this data set, in which bias has been deliberately manipulated by including black vocabulary items, is to determine the validity of the ANOVA method when the biased items are known externally. Specially, these data consists of responses to fifty multiple-choice vocabulary test items, including ten black slang items which were intentionally written by a black author to be biased against whites, independent of any statistical index of item bias. The other forty items were drawn from the verbal section of the College Qualification Test which is an aptitude test constructed for college students. There were 1,022 whites and 1,008 blacks. A range of total scores for whites was 11 to 48 and that for blacks was 7 to 48. Further details of the data are provided by the Subko''iak et al's study. (1984).

For this study, matched-pair 60 subjects whose total test scores were from 11 to 48 were selected randomly from all whites and blacks in order to determine the ANOVA method with small matched two group data.

Results and Conclusion

The ANOVA table for the matched-pair data were reported in Table 1. Obviously, there was no significant group ability difference. The significant interaction between the group and the item existed. Item difficulty on each item for each group, group abilities, interaction contrasts, a signed measure of an item bias index, and an unsigned measure were reported in Appendix A. None of easy or difficult item except the black slang items was detected as bias. This study imbedded the item bias indices of the Rudner's area measure from the Subkoviak et. al's study (1984) because the Rudner's area measure procedure is the most sound method to detect biased items. The resulting item bias indices for the ANOVA method and the Rudner's area measures were reported in Appendix B.

insert Table 1 about here

Correlation Pearson correlations between the á priori bias index (zero-one coding) and the bias index for the matched-pair data was shown in Table 2. The signed measures of the ANOVA method were highly correlated with the á priori bias index ($r >$ .819). This result was better than the result of the Seong and Subkoviak's study (1987). In their study, the signed measure of the Linn and Harnisch's pseudo-IRT(Z) method was correlated .762 with the á priori bias index, the Camilli's chi-square was .798, and the Angoff's revised was .522.

insert Table 2 about here

Percentage Agreement The signed measure with the matched-pair 60 subjects from each group detected seven black slang items out of ten black slang items as an agreement rate of 94%. The unsigned measure achieved 90 %. Agreement rates were shown in Table 3.

insert Table 3 about here

Agreement Among Methods Intercorrelation among the bias indices of the Rudner's area measure and the bias indices of the ANOVA method with the matched-pair subjects were reported in Table 4. The ANOVA procedure analyzing the matched-pair subjects was correlated highly with the Rudner's area measure method for both signed and unsigned measures (r=.890, .812).

insert Table 4 about here

This study showed that the ANOVA method for detecting item bias was applicable when there was no significant difference between group abilities and there was a small number of students within groups.

Hunter's criticism (1975) for the ANOVA method is sound theoretically. Easy or difficult items tend to be detected as bias when group abilities differ. However, in this study, none of the easy and difficult items except the black slang items was detected as a biased item because of matching group abilities (See Appendix A). In other word, there was no group ability difference to misjudge easy or difficult items as biased items. Another reason may be that the non significant group effect is orthogonal to the interaction effect between item and group when test bias exists. This study confirmed Jensen's assertion (1980) that significant interaction still exists in an analysis of variance with the matched-pair data if test bias exists in a test. This study concludes that ANOVA approach with small data is still valid to detect biased items after matching one group's ability and another group's ability. Group ability may be contaminated to the extent that biased items are included in the computation of ability estimates. This study suggests that the analysis of variance is repeated to get the second error term after excluding a biased item and the group abilities are recalculated. The amount of bias measured by the ANOVA method is still more a function of the ability location than all range of group ability.

Reference

CAMILLI,G. (1979). A critique of the chi-square method assessing item bias. Unpublished paper, Laboratory of Educational Research, University of Colorado, Boulder.

CAMILLI,G. & SHEPARD,L. (1987). The inadequacy of ANOVA for detecting test bias. Journal of Educational Statistics. 12, 87-99.

CLEARY,T.A., & HILTON,T.L. (1968). An investigation of item bias Educational and Psychological Measurement, 28, 61-75

HOLLAND,P.W., & THAYER,D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.). Test Validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

HUNTER,J.E. (1975). A critical analysis of the use of item means and item-test correlation to determine the presence or absence of content bias in achievement test items. Paper presented at the National Institute of Education Conference on Test Bias, Annapolis, MD.

JENSEN,A.R. (1980). Bias mental testing. New York: Free Press,

LINN,R.L., & HARNISCH,D.L. (1981). Interaction between item content and group membership on achievement test items. Journal of Educational Measurement, 18, 109-118.

LORD,F.M. (1977). A study of item bias using item characteristic curve theory. In Y.H.Poortinga (Ed.), Basic problems in cross-cultural psychology (pp. 19-29). Amsterdam: Swets and Zeitlinger.

RUDNER,L.M. (1977). An evaluation of select approaches for biased item identification. Unpublished doctoral dissertation, Catholic University of America, Washington, D.C.

SEONG,T.J & SUBKOVIAK,M.J. (1987). The comparative study of recently proposed item bias detection method. Paper presented at the annual meeting of NCME., Washington DC. April. (ERIC Document Reproduction Service NO. ED281883)

SHEPARD,L., CAMILLI,G. & WILLIAMS,D.M. (1985). Validity of approximation techniques for detecting item bias. Journal of Educational Measurement, 22, 77-105

SUBKOVIAK,M.J., MACK,J.S., IRONSON,G.H., & CRAIG,R.D.(1984). Empirical comparison of selected item detection procedures with bias manipulation. Journal of Educational Measurement, 21, 49-58.

Table 1 Analysis of Variance Table for Matched-Pair Data (60 Whites and 60 Blacks)

SOURCE	DF	SS	MS	F
Between block (student)				
Group	1	.004	.004	.00
Block w.a group	118	172.262	1.460	I Error
Within block (student)				
Item	49	265.005	5.408	32.22
Group x Item	49	80.291	1.651	9.84
Item x Block w.a Group	5782	970.454	.168	II Error
Total	5999	1488.016		

Table 2 Correlations between the a Priori Bias Index and Detected Bias Measure (T) with Matched-Pair Data Sets

Data Set	Signed Measure	Unsigned Measure
Matched 60	.819	.681

Table 3. Contingency Table and Percentage Agreement of the ANOVA Method with Matched-Pair Data Sets for Detecting the Imbedded Bias in the Test

Data	A priori / Detection	Signed Measure			Unsigned measure		
		Bias	Unbias	%	Bias	Unbias	%
Matched 60	Bias	7	3	94	5	5	90
	Unbias	0	40		0	40	

Table 4. Intercorrelations between the Rudner Area Bias Measure and the ANOVA Bias Measures with the Matched-Pair Data

	Rudner's ICC Area Measure	
	Signed Measure	Unsigned Measure
ANOVA	.890	.812

Appendix A Average Score, Ability, Interaction Contrast, Signed Measure and Unsigned Measure of ANOVA Method with the Matched-Pair Data (60 Whites and 60 Blacks)

Item	Item Difficulty		Group Ability		Contrast	S(T)	US(T)
	White	Black	White	Black			
1E	0.80000	0.70000	0.539116	0.539455	-0.100339	-1.32794	1.32794
2	0.71667	0.60000	0.540816	0.541496	-0.117350	-1.55307	1.55307
*3	0.51667	0.86667	0.544898	0.536054	0.358844	4.74913	4.74913
4	0.55000	0.40000	0.544218	0.545578	-0.151360	-2.00317	2.00317
5E	0.91667	0.85000	0.536735	0.536394	-0.066329	-0.87783	0.87783
6	0.71667	0.78333	0.540816	0.537754	0.069722	0.92273	0.92273
*7	0.08333	0.21667	0.553742	0.549319	0.137763	1.82323	1.82323
8E	0.93333	0.80000	0.536395	0.537414	-0.134349	-1.77805	1.77805
9D	0.18333	0.31667	0.551701	0.547278	0.137763	1.82322	1.82322
10	0.61667	0.63330	0.542857	0.540816	0.018671	0.24710	0.24710
11	0.68333	0.56667	0.541497	0.542176	-0.117339	-1.55293	1.55293
12	0.70000	0.60000	0.541156	0.541496	-0.100340	-1.32795	1.32795
13	0.73333	0.46667	0.540476	0.544217	-0.270401	-3.57863	3.57863
14	0.83333	0.60000	0.538435	0.541496	-0.236391	-3.12852	3.12852
*15	0.51667	0.73333	0.544898	0.538775	0.222783	2.94843	2.94843
16	0.70000	0.61667	0.541156	0.541156	-0.083330	-1.10283	1.10283
17D	0.11667	0.20000	0.553061	0.549659	0.086732	1.14785	1.14785
18	0.33333	0.30000	0.548639	0.547618	-0.032309	-0.42760	0.42760
19	0.31667	0.23333	0.548980	0.548979	-0.083339	-1.10295	1.10295
*20	0.51667	0.98333	0.544898	0.533673	0.477885	6.32458	6.32458
21	0.71667	0.56667	0.540816	0.542176	-0.151360	-2.00318	2.00318
22D	0.30000	0.18333	0.549320	0.549999	-0.117349	-1.55306	1.55306
23E	0.85000	0.88333	0.538095	0.535714	0.035711	0.47262	0.47262
*24	0.41667	0.88333	0.546939	0.535714	0.477885	6.32458	6.32458
25	0.83333	0.58333	0.538435	0.541836	-0.253401	-3.35364	3.35364
26	0.38333	0.38333	0.547619	0.545918	0.001701	0.02251	0.02251
*27	0.40000	0.80000	0.547279	0.537414	0.409865	5.42436	5.42436
28	0.50000	0.36667	0.545238	0.546258	-0.134350	-1.77805	1.77805
29	0.48333	0.25000	0.545578	0.548639	-0.236391	-3.12852	3.12852
30	0.41667	0.25000	0.546939	0.548639	-0.168370	-2.22829	2.22829
31E	0.85000	0.75000	0.538095	0.538435	-0.100340	-1.32795	1.32795
32	0.70000	0.60000	0.541156	0.541496	-0.100340	-1.32795	1.32795
33	0.70000	0.53333	0.541156	0.542857	-0.168370	-2.22830	2.22830
*34D	0.00000	0.18333	0.555442	0.549999	0.188773	2.49831	2.49831
35E	0.80000	0.70000	0.539116	0.539455	-0.100339	-1.32794	1.32794
36	0.70000	0.35000	0.541156	0.540598	-0.355442	-4.70410	4.70410
*37	0.18333	0.88333	0.551701	0.535714	0.715987	9.47575	9.47575
38	0.80000	0.56667	0.539116	0.542176	-0.236390	-3.12851	3.12851
39	0.30000	0.46667	0.549320	0.544217	0.171773	2.27333	2.27333
40	0.23333	0.25000	0.550680	0.548639	0.018711	0.24763	0.24763
41	0.35000	0.26667	0.548299	0.548299	-0.083330	-1.10283	1.10283
42E	0.95000	0.93333	0.536054	0.534693	-0.015309	-0.20261	0.20261
*43	0.53333	0.76667	0.544558	0.538095	0.239803	3.17368	3.17368
44	0.40000	0.50000	0.547279	0.543537	0.103742	1.37298	1.37298
45	0.50000	0.40000	0.545238	0.545578	-0.100340	-1.32795	1.32795
46	0.68333	0.56667	0.541497	0.542176	-0.117339	-1.55293	1.55293
47D	0.30000	0.13333	0.549320	0.551020	-0.168370	-2.22829	2.22829
48	0.43333	0.33333	0.546599	0.546938	-0.100339	-1.32794	1.32794
49	0.85000	0.51667	0.538095	0.543197	-0.338432	-4.47898	4.47898
*50	0.16667	0.81667	0.552041	0.537074	0.664967	8.80052	8.80052

* Black Slang Item E Easy Item D Difficult Item

Appendix B   Item Bias Indices

Item	A Prior	Signed ICC-3	Unsigned ICC-3	Signed T	Unsigned T
1	0	0.06	0.96	-1.32794	1.32794
2	0	0.08	0.45	-1.55307	1.55307
*3	1	3.04	3.65	4.74913	4.74913
4	0	0.27	0.27	-2.00317	2.00317
5	0	0.05	0.34	-0.87783	0.87783
6	0	0.78	0.92	0.92273	0.92273
*7	1	2.05	2.12	1.82323	1.82323
8	0	-0.78	0.90	-1.77805	1.77805
9	0	0.77	0.77	1.82322	1.82322
10	0	0.08	0.08	0.24710	0.24710
11	0	0.24	0.24	-1.55293	1.55293
12	0	0.15	0.54	-1.32795	1.32795
13	0	0.07	0.37	-3.57863	3.57863
14	0	0.04	0.46	-3.12852	3.12852
*15	1	1.32	1.32	2.94843	2.94843
16	0	0.49	0.47	-1.10283	1.10283
17	0	0.70	0.82	1.14785	1.14785
18	0	-0.05	0.56	-0.42760	0.42760
19	0	0.05	0.10	-0.10295	0.10295
*20	1	4.64	4.69	6.32458	6.32458
21	0	-0.34	0.49	-2.00318	2.00318
22	0	0.02	0.15	-1.55306	1.55306
23	0	0.10	0.11	0.47262	0.47262
*24	1	7.25	7.25	6.32458	6.32458
25	0	-0.39	0.43	-3.35364	3.35364
26	0	0.44	0.44	0.02251	0.02251
*27	1	3.42	3.77	5.42436	5.42436
28	0	-0.16	0.44	-1.77805	1.77805
29	0	-0.63	0.77	-3.12852	3.12852
30	0	0.20	0.49	-2.22829	2.22829
31	0	0.34	0.34	-1.32795	1.32795
32	0	-0.29	0.54	-1.32795	1.32795
33	0	0.34	0.37	-2.22830	2.22830
*34	1	3.34	3.37	2.49831	2.49831
35	0	0.05	0.33	-1.32794	1.32794
36	0	-0.14	0.22	-4.70410	4.70410
*37	1	6.36	6.36	9.47575	9.47575
38	0	0.45	0.61	-3.12851	3.12851
39	0	0.45	0.96	2.27333	2.27333
40	0	0.03	0.55	0.24763	0.24763
41	0	-0.82	0.82	-1.10283	1.10283
42	0	-0.27	0.27	-0.20261	0.20261
*43	1	3.33	3.33	3.17368	3.17368
44	0	0.68	0.68	1.37298	1.37298
45	0	0.10	0.23	-1.32795	1.32795
46	0	0.29	0.64	-1.55293	1.55293
47	0	-0.95	0.95	-2.22829	2.22829
48	0	0.20	0.20	-1.32794	1.32794
49	0	-0.36	0.36	-4.47898	4.47898
*50	1	5.61	5.61	8.80052	8.80052

* Black Slang Item