Volume Title: Higher Education and Earnings: College as an Investment and Screening Device

Volume Author/Editor: Paul J. Taubman, Terence Wales

Volume Publisher: NBER

Volume ISBN: 0-07-010121-3

Volume URL: http://www.nber.org/books/taub74-1

Publication Date: 1974

Chapter Title: Appendix E: Response Bias, Test Scores, and Factor-Analysis

Chapter Author: Paul J. Taubman, Terence Wales

Chapter URL: http://www.nber.org/chapters/c3660

Chapter pages in book: (p. 197 - 210)

Appendix E: Response Bias, Test Scores, and Factor-Analysis

The usefulness of any sample depends in part on the accuracy of its information. In this appendix we explore this issue, discuss the construction of the ability measures, and examine the 1969 respondents to determine the extent of "response" and success biases. We begin our discussion by comparing the educational responses of 1955 and 1969. It is useful to compare the accuracy of the two answers as an indication of the accuracy both of the respondent's memory and of the education data. Also, it is of some interest to observe how much additional education was undertaken a decade after the end of World War II.

In each row in Table E-1 we present the distribution of 1969 educational responses (by the 1955 response) for each person who answered the questionnaire in 1969 and for whom we had a 1955 response.[1] Several features of the table are immediately apparent. First, the largest element in each row always occurs in the 1969 category that corresponds to the 1955 answer, that is, on the diagonal. For all but the high school and some-graduate-work groups, over 70 percent of the people gave the same answer in both years.

Second, most of the differences between the 1969 and 1955 answers indicate higher educational achievement in 1969. This difference could be expected for two reasons other than response error. First, in the 1955 sample, those who completed their education before 1946 were counted as having only a high school education, while in 1969 their correct education could be given.[2] Second, individuals could have achieved more education after 1955.

[1]Unfortunately, some of the 1955 questionnaires had been misplaced and were not available to us.

[2]This would, of course, be a measurement error in the 1955 data.

TABLE E-1 Comparison of education reported for 1955 and 1969 by those responding in 1969

Education as reported in 1955*	Education as reported in 1969					
	High school	Some college	Undergraduate degree	Some graduate work†	Master's‡	Ph.D.§
High school	.57	.34	.078	.005	.007	.003
•	842	500	116	8	10	5
Some college	.14	.71	.10	.03	.01	.01
	88	463	65	21	9	7
Undergraduate degree	.01	.05	.79	.09	.05	.01
	15	57	941	107	61	10
Some graduate work†	.00	.09	.19	.40	.24	.08
	0	19	38	81	48	16
Master's‡	.01	.01	.04	.02	.78	.15
	2	4	11	6	242	45
Ph.D.§	.00	.01	.02	.01	.02	.94
	0	1	4	3	6	229

*Excludes those from whom education was not reported in 1955.

†Includes those with some graduate course work but no degree.

‡Includes those with an LL.B. or a J.D. and those with more credit than a master's but less than a Ph.D.

§Includes all graduates from a graduate program of three years or more.

NOTE: The upper entry in each cell is the percentage of the row sum. The lower entry is the number of people in the cell.

To determine what percentage of the greater educational achievement reported in 1969 is due respectively to missing information in 1955, to response errors, and to post-1955 education, we separated the items in each cell to the right of the diagonal into three groups of education completed: prior to 1946, from 1946 through 1955, and after 1955.[3] The results, which are given in Table E-2, indicate quite clearly that the response error is rather small. That is, approximately 80 percent of all individuals counted as high school graduates in 1955 but with higher reported education in 1969 can be accounted for by the lack of

[3] These categories are not quite perfect, since, for example, "some college" completed after 1955 could have been begun before 1955; and since a Ph.D. could be completed after 1955, the 1955 response in Thorndike and Hagen could still be in error.

TABLE E-2
Distribution of those with more education reported in 1969 than in 1955, by date of last year attended school

Education as reported in:		Last year attended school		
1955	1969	Before 1946	1946–1955	After 1955
High school	Some college	.828	.078	.094
		414	39	47
	Undergraduate degree	.759	.129	.112
		88	15	13
	Some graduate work	.500	.000	.500
		4	0	4
	Master's	.200	.100	.700
		2	1	7
	Ph.D.	.400	.200	.400
		2	1	2
Some college	Undergraduate degree	.138	.492	.369
		9	32	24
	Some graduate work	.095	.381	.524
		2	8	11
	Master's	.000	.111	.889
		0	1	8
	Ph.D.	.143	.571	.286
		1	4	2
Undergraduate degree	Some graduate work	.000	.477	.523
		0	51	56
	Master's	.016	.098	.885
		1	6	54
	Ph.D.	.000	.200	.800
		0	2	8
Some graduate work	Master's	.000	.229	.771
		0	11	37
	Ph.D.	.000	.063	.938
		0	1	15
Master's	Ph.D.	.000	.089	.911
		0	4	41

pre-1946 data in the 1955 questionnaire.[4] An additional 12 percent had continued their education after 1955 (with most in the some-college group). Thus, only about 6 percent of the 973 individuals who reported a high school education in 1955 made a response error.[5] Much of the extra education in 1969 in the other 1955 groups also represents post-1955 achievement.

After eliminating all cases of post-1955 achievement, we find that the response errors as a percentage of all persons reporting a given education in 1955 are:

• High school	6
• Some college	20
• Undergraduate degree	11
• Some graduate work	35
• Master's	9
• Ph.D.	6

Sizable response errors occur only in the some-college and some-graduate-work cases. In these instances, a large number of people who reported some college or postgraduate work, respectively, in 1955 did not report this education in 1969. We suspect that these people had a minimal amount of college or graduate work and reported it in 1955 either because they anticipated continuing their education or because the short stay was recent enough to remember. By 1969, neither of these reasons would be operative. If this logic is correct, such response bias would not be important either in terms of the reliability of memory or in the computation of regressions, though some recent work by A. Haspel has indicated that slightly higher $\bar{R}^2$s can be obtained by substituting the 1969 response when there is a response error. Overall, we would judge the response error for educational achievement to be very small in the sample.[6]

[4]It is also encouraging to note that only 13 people with a 1955 education greater than high school and a 1969 educational level above that of 1955 reported completing their education before 1946. Such early education should not have been reported in the 1955 data.

[5]In calculating the response rate, we have not used the 508 people who had completed their education before World War II because the 1955 questionnaire did not ask for this information.

[6]Some of the differences between the 1955 and 1969 responses reflect coding and punching errors and not a response error.

However, the misclassification arising from the lack of pre-1946 education in the 1955 sample is important, and as shown in Appendix F, this misclassification significantly alters our estimates of the effect of education on income.

It is interesting to note how many people in the sample continued their education after 1955. Table E-2 indicates that 329 people advanced to a new educational plateau after 1955.[7] Of these, 47 achieved some college education, 37 an undergraduate degree, 71 some postgraduate training, 106 an M.A. or equivalent, and 68 a Ph.D. or equivalent.

In the above discussion, and in the analysis in our study, a person is considered to have had a college education if he listed any education other than vocational after high school. The coding of the names of schools was finished after this study was completed. As should be expected, the quality of the schools varies greatly.[8] Not only are there Princetons and Podunks but there are also a few instances of unaccredited schools and several mortuary and embalming colleges. The effect of different institutions on earnings is the subject of a separate study already under way and will be explored here only briefly.

The people in our sample are much more highly educated than veterans of World War II in general. For example, of those veterans aged 25 through 34 in 1952 who had at least a high school diploma, about 60 percent had a high school education only, 20 percent had some college, and 20 percent had at least one college degree.[9] On the other hand, only one-quarter of our sample had a high school education only, one-quarter had some college, and the rest had graduated with at least one degree (see Table E-3.)

We can also consider briefly the accuracy of the earnings and ability data. The individual tests are fairly reliable in the sense that comparable results are obtained when identical or similar tests are administered to the same people after the passage of time. For several reasons, we judge the earnings data reported

[7]Of course, others may have continued their education without changing education categories.

[8]By "quality" we mean ranking on the Gourman Index, which, like the College Boards, has a mean of 500 and a standard deviation of 100.

[9]See, for example, Miller (1960, p. 997). Although some of these people would have attained more education by 1955, the differences between those in our sample and World War II veterans in general would still be large. Moreover, no one in our sample had less than a high school diploma.

TABLE E-3
Distribution of education for the 1955 and 1969 samples

	High school	*Some college*	*Under-graduate degree*	*Some graduate work*	*Master's*	*Ph.D., LL.B., and M.D.*
(1)	.416	.165	.262	.043	.060	.054
(2)	.363	.160	.292	.050	.076	.060
(3)	.232	.256	.288	.054	.092	.076

NOTE: Row 1 is based on the entire 1955 sample except for those for whom there was no information on education.

Row 2 is based on those who responded in 1969, using the education reported in 1955.

Row 3 is based on those who responded in 1969, using the education reported in 1969.

in 1969 and 1955 for those years, respectively, to be reliable—at least in comparison with other studies. First, such items as interest and dividends, which are more irregular in amount and timing, do not have to be recalled. Second, except for six people whose earnings exceeded $100,000 in 1969 and for whom $99,999 was used, we recorded the actual dollar amounts, and do not have to use midpoints of classes or estimate the mean of open-ended classes. As customary in such samples, however, large numbers of people reported such round figures as $10,000. Third, we have already shown that the education response is accurate.

We next consider whether there is a bias in the response rate of the variously educated groups. To do this, we calculate the distribution by education for all those in the 1955 sample and for those who responded in 1969. Since we do not know the pre-1946 education of those who did not respond in 1969, we use the educational information as recorded in 1955 for this comparison.

Comparing the first two rows of Table E-3, it is clear that in 1969 there was a response bias, with the more educated tending to be more likely to respond, though this bias does not appear to be very large. In row 3 we have the actual education of the 1969 respondents. The difference between rows 2 and 3, which reflects primarily the education obtained before 1946 and after 1955, confirms the results in Table E-1.

THE TEST SCORES AND FACTOR ANALYSIS

We have scores on 17 different tests in the sample. A brief description of the tests, taken from Thorndike and Hagen (1959, pp. 9–12), indicates the types of material and abilities covered in each test.[10]

Reading Comprehension A test of comprehension of rather technical passages. Primarily a measure of verbal comprehension, it also depended somewhat on mechanical experience and on general reasoning ability.

General Information–Navigator An information or vocabulary test dealing with the terminology of such fields as astronomy, trigonometry, and science. The common factor measured was one of verbal ability, though the content suggests that the test also reflects technical and quantitative academic leanings.

Mathematics A test of knowledge of advanced arithmetic, algebra in large amount, and trigonometry to some extent. Performance appeared to depend upon a complex of factors, including numerical fluency, verbal comprehension, background in school mathematics, and general reasoning ability. [This is referred to in the text as Mathematics A.]

Arithmetic Reasoning A test composed of arithmetic word-problems of the sort that are common in the subtests of both scholastic aptitude and school achievement batteries, but cast in terms of planes and aviation. Numerical fluency, general reasoning, and verbal comprehension were the main factors identified in this test. [This is referred to in the text as Mathematics B.]

Numerical Operations I and II These were two scores, obtained separately for scoring convenience, representing speed and accuracy in simple numerical computations. Test I included addition and multiplication, Test II subtraction and division. These were nearly pure tests of numerical fluency.

Dial and Table Reading The first part of this test required the subject to determine the numerical value represented by the reading on one of a set of airplane instrument dials, whereas the second part required him to locate entries in rather complex numerical tables. The test involved a number of ability factors, primarily numerical fluency, spatial relations, and perceptual speed.

[10]For the convenience of the reader, we also include the description of the information in the biographical items. The pilot and navigator biography keys are *not* included in the factor analysis.

Speed of Identification This was a speeded test requiring the matching of visual forms. It was a rather pure measure of speed of visual perception.

Spatial Orientation I This was also a perceptual matching test. In this case, small excerpts from aerial photographs had to be matched against locations in the complete photographs. Perceptual speed was again the chief factor involved.

Spatial Orientation II This test required the matching of sections of aerial photographs with the locations on aerial maps corresponding to the photographs. Perceptual speed was again a major factor, together with some dependence upon spatial visualizing and on social studies background.

Mechanical Principles This test was patterned closely after Benett's Mechanical Principles Test, except that the items were cast in an aviation setting. Pictorial situations presented problems involving mechanical concepts and principles. Performance on this test depended primarily on factors of mechanical experience and of visualizing ability, but also to a slight degree on verbal comprehension and general reasoning.

Two-Hand Coordination This was an individual apparatus test in which the examinee, by manipulating lathe-type controls, had to keep a pointer in contact with an irregularly moving target button. Performance on the test appeared to depend in part upon mechanical experience, in part upon spatial factors, and in part upon psychomotor coordination.

Biographical Data Blank—Pilot Key The Biographical Data Blank was an assembly of questions about home and family background, school activities and successes, out-of-school activities and hobbies, work experience, and certain other items. The pilot key was an empirical keying of those items that were found to predict success in pilot training. The main factor represented in the pilot key was mechanical experience.

Complex Coordination This test was an individual apparatus one in which the examinee manipulated stick and rudder controls similar to those encountered in a light airplane. By moving the controls appropriately, he could match patterns of stimulus lights presented to him. His task was to match as many of the stimulus patterns as he could in a limited test period. The test appeared to measure primarily factors of spatial relations and psychomotor coordination.

Rotary Pursuit with Divided Attention The subject's basic task was to keep a metal stylus resting on a target button that rotated on a 78 rpm phonograph turntable. At the same time, with the other hand, he had to depress the response button corresponding to one of two signal lights. The test appeared to measure psychomotor coordination as well as some type of spatial factor.

Finger Dexterity This was a peg-turning test in which the subject was required to remove each of a series of square pegs from its hole, rotate it through 180°, and reinsert it. He was scored in terms of the number of pegs turned in a series of brief trials. Common factors measured by the test included psychomotor coordination and precision, visualizing, and possibly a little perceptual speed.

Aiming Stress Basically, this was a test of hand steadiness, in which the subject had to keep a stylus pointer in a target hole without touching the edges of the hole. Verbal heckling was introduced in some trials to try to add an element of emotional stress. However, the only common factor appearing in this test was one of psychomotor coordination, so it should be thought of primarily as a motor test.

General Information—Pilot This test was made up of a rather heterogeneous collection of information items selected because they appeared to have validity for predicting success in pilot training. These included items about planes and flying, auto driving, sports, and hobbies. The test appeared quite heterogeneous, and involved factors of pilot interest, verbal comprehension, mechanical experience, and perceptual speed.

Discrimination Reaction Time This test was an individual test of speed of reaction. The reaction was to a pattern of visual stimuli, and the subject was required to choose the correct one of four response switches. Stimulus patterns were presented one after another, at about four second intervals, and the subject had to respond as rapidly as possible to each as it occurred. The test measured a complex of different factors, including spatial relations, psychomotor precision, perceptual speed, and visualizing.

Biographical Data Blank—Navigator Key A second empirical key was established for the Biographical Data Blank, weighting those items that were found to predict success in navigator training. The key seemed to provide primarily a measure of extent of mathematical background.

From the titles and the above descriptions, it is clear that many of the tests measure different facets of the same ability. Rather than use all the tests or any arbitrary subset, we decided to use factor analysis to obtain a measure of a few types of ability.[11] The actual factor analysis was conducted by A. Beaton.

The basic idea in factor analysis is that any test contains information on one or more general abilities and a test-specific content. That is,

$$SC = F + u$$

where *S* is the set of scores; *F* represents the set of general abilities; and *u* represents the test-specific components. Using the scores on each of the tests, it is possible to estimate C by imposing certain conditions on *u*. Estimates of *F* can then be found from *SC* where C is the factor loadings. Each *F* is, of course, just a weighted average of the test scores.[12] In some instances, however, the major weights in each average are attached to items that measure one type of attribute; this attribute is then used to describe the factor.

In the NBER-TH data, the factor loadings for each of four factors are as in Table E-4. Consider the second factor, in which Rotary Pursuit, Two-Hand Coordination, and Complex Coordination all have loadings in excess of .65. In addition, Finger Dexterity, Aiming Stress, Discrimination Reaction Time, and Mechanical Principles have weights in the range .35 to .54. Since the common element in all tests is coordination, we refer to this as the complex-coordination factor. For the fourth factor, the only important tests are Speed of Identification and the two Spatial Orientation tests. Given the description of the Speed of Identification test, it seems clear that the fourth factor measures spatial perception and, perhaps, abstract reasoning.

Both these factors are easy to interpret or identify. The first and third are somewhat more difficult, although the first is clearer. In the first factor, the most important items, with loadings of at least .69, are Numerical Operations, mathematics (Math A), arithmetic reasoning (Math B), and Dial and Table

[11]It is, of course, true that the factors are linear combinations of the test scores; thus the same results could be achieved by entering linearly all the scores in our regressions. However, we are interested in studying nonlinear effects of various ability measures and interactions of ability with education, as well as determining which types of ability influence earnings. For these purposes it is appropriate to use the factors.

[12]Because the original test scores are standardized and then manipulated as correlations, all weights have to lie between +1 and −1. The importance of each test in a factor is indicated by the absolute size of its loading coefficient.

TABLE E-4
Factor loadings

Ability test	*Factor loadings* 1	2	3	4
Reading Comprehension	0.4123	0.0700	0.7186	0.0136
Mechanical Principles	0.0149	0.3522	0.7210	0.0247
Dial and Table Reading	0.6990	0.2566	−0.0129	0.3260
Spatial Orientation II	0.0658	0.1042	0.3117	0.6420
Spatial Orientation I	0.2217	0.1379	0.0311	0.7642
Numerical Operations I and II	0.7822	0.0597	−0.2183	0.1030
Speed of Identification	0.0500	0.1008	0.0643	0.7831
General Information–Navigator	0.4842	−0.1199	0.5605	0.1495
General Information–Pilot	−0.0811	0.0329	0.5874	0.3567
Mathematics B	0.7444	−0.0104	0.3514	−0.0717
Mathematics A	0.7464	−0.0469	0.3060	0.5071
Rotary Pursuit	−0.0304	0.6772	0.0453	0.0396
Two-Hand Coordination	−0.0385	0.6870	0.2703	0.0572
Complex Coordination	0.1251	0.7026	0.1877	0.2028
Aiming Stress	−0.0111	0.4128	0.0009	0.0093
Discrimination Reaction Time	0.3891	0.3800	0.0940	0.2636
Finger Dexterity	0.1974	0.5438	−0.1664	0.1253

Reading. All these tests are concerned with mathematics and quantitative skills and primarily measure numerical fluency or speed and accuracy, but not necessarily problem-solving techniques. Secondary but still important weights (.48 to .39) are accorded to General Information–Navigator, Reading Comprehension, and Discrimination Reaction Time. The navigator test does emphasize mathematical material, but the other two items do not. Nevertheless, we treat this as a mathematical-ability test.

In the third factor, Reading Comprehension and Mechanical Principles have loadings in excess of .7. General Information–Pilot and General Information–Navigator have loadings of about .5, while Math B and A have loadings of .35 and .30, and Spatial Orientation II has a loading of .31. In general, these tests encompass verbal ability, mathematical skills, reasoning, *and* *mechanical principles*. The first three items would be found in standard verbal-IQ tests, but mechanical principles would not. We have chosen to call this third factor IQ.[13] It is important to note, however, that Thorndike has written us to say he believes that the first factor would correlate much more closely with an IQ measure and should be named as such, while the third factor tends toward mechanical principles. The reader should keep this caveat in mind when examining our remarks about the importance of the different types of ability and when comparing the NBER-TH results with those of Wolfle-Smith.

The tests were administered in 1943, whereas the primary earnings data are for 1955 and 1969. If tested in 1955 or 1969, a person would not score the same on each test (or achieve the same standing on a particular factor) because of post-1943 education, general maturation effects, and sampling error from any test. The test-retest validity on the tests used was fairly high, so we can ignore the last possibility. The effect of post-1943 education on particular cognitive skills can be captured by, or impounded in, the overall effect of education on earnings. The available evidence in, for example, Bloom (1964) would indicate that, on the average, maturation effects are small between the ages of 20 to 50, though this may not be true for soldiers in a war. However, in our analysis we only use a dummy variable indicating to which fifth of the distribution a person belonged. Thus, we merely need to assume that a person's relative ability position did not change enough over time to move him to a different fifth.

Finally, it might be asked if the differences in pretest education affect the test scores and, thus, the ability measure. In Chapter 5, we demonstrated that the differences in the quantity and regional quality of pretest education had little or no effect on the scores.

It is of some interest to study the response bias in terms of ability and biography measures.[14] If the 3,743 respondents in 1969 whom we study were a random drawing, there would be

[13] Of course, we have constructed the factors to be orthogonal (for the 9,700 people). Although IQ and mathematical tests in general are correlated, it would be possible to factor-analyze such tests to extract several factors.

[14] As discussed below, the biography variable represents a mixture of ability, education, and socioeconomic background factors.

about 748 people in each fifth.[15] The number of 1969 respondents in each fifth for the various factors, which are presented in Table E-5, clearly indicate a tendency for respondents to be in the higher ability ranges. This is very pronounced for the mathematical and IQ factors, but it is also true for the other two.

SUCCESS BIAS

We turn now to the question of whether there is a success bias in the 1969 respondents. We assume that success persists over time, so that the successful in 1955 would also be successful in 1969, and vice versa.[16] The success bias implies that those with above-average earnings will be more likely to respond. This bias cannot be measured simply by comparing average income of 1969 nonrespondents with that of the 1969 respondents, since we know from the preceding analysis that the 1969 respondents have, on the average, more education and ability. But the existence of a success bias would imply that the equation for earnings in 1955 for the nonrespondents would be different from the corresponding equation for respondents. The details of these regressions are presented in Appendix F but we briefly summarize the important findings here. Holding various ability measures, education, and time on the job constant, we do not reject the null hypothesis that the respondents and nonrespondents are randomly drawn from the same population. Based on individual coefficients, the only category in which there may be

[15]Some of the initial 4,443 observations were eliminated because of missing data. The quintile points were calculated from the distribution of the 9,710 individuals in the 1955 sample.

[16]If Mincer's on-the-job-training argument is correct, the validity of this assumption is weakened, though not invalidated.

TABLE E-5
Number of 1969 respondents in 1955-sample fifths for four ability factors

	Fifths (low to high ability)				
	1	2	3	4	5
Factor 1 (mathematical)	640	688	737	803	875
Factor 2 (complex coordination)	724	724	732	775	788
Factor 3 (IQ)	672	704	738	776	853
Factor 4 (spatial)	694	688	763	780	818
Biography variable	732	732	702	747	830

a bias is the three-year graduate degree (Ph.D., LL.B., and M.D.). In this category, the less successful individuals appear to have responded more.

The tests for success bias and the examination of the effects of measurement error lead to the following conclusions. The 1969 respondents do not contain an overrepresentation of the successful in each of the education-ability cells; hence, leaving aside the advantages of a larger sample size, the 1969 respondents could be used to represent the whole Thorndike group. The error introduced by not using the correct pre-1946 education data, however, has serious consequences for studying the returns to education. Thus, unless the advantages of the larger sample outweigh the disadvantages of the measurement error, we should use the 1969 respondents. For most of our purposes, a sample of about 4,000 is nearly as good as one of 8,000. In addition, we have collected other data in 1969 that are important determinants of income. For these reasons we use the smaller but more accurate sample in our analysis.