DOCUMENT RESUME

ED 103 462 TM 004 295

AUTHOR Findley, Warren G.
TITLE Language: Friend or Foe?
PUB DATE Apr 74
NOTE 12p.; Discussion of papers at NCME symposium on "The International Educational Achievement Study", Methodological Issues and Selected Results (Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Academic Achievement; *Comparative Education; *Cultural Factors; Culture Free Tests; Educational Accountability; Educational Assessment; Foreign Students; Language; Language Handicaps; Multiple Choice Tests; Reading Comprehension; Test Bias; Test Construction; *Testing; *Testing Problems
IDENTIFIERS International Evaluation Educational Achievement

ABSTRACT

The role of testing has received a great deal of criticism over the years. Some steps have been taken to develop more universal items into the test. One innovation was the introduction of the multiple choice exam. Yet even with this type of aptitude test, the majority of questions and plausible answers have become quite involved and intense. All this requires a great deal of concentration for the reader to comprehend the question as well as the four to five plausible answers all within a certain amount of time. The persons who risk the most in these situations are those to whom the language of the test is a secondary one. By citing some of the research findings that were carried out by members of NCME, the author tries to achieve an analytical and interpretative reply to these findings. The author also makes some proposals which might remedy the errors of the testing instrument especially within the area of international studies. (DEP)

Language: Friend or Foe?*

Warren G. Findley
University of Georgia

In participating as discussant in this symposium on the International Educational Achievement study, it gives me great satisfaction to look back twelve years to a meeting of the Advisory Committee of the USOE Cooperative Research Program where nine of us voted to recommend that the proposal for this study be funded initially. That vote of confidence has been amply justified over the years by work and publications, culminating for the moment in three volumes on Science Education in Nineteen Countries, Literature Education in Ten Countries, and Reading Comprehension in Fifteen Countries. These were, of course, preceded in 1967 by a two-volume International Study of Achievement in Mathematics (in twelve countries).

Some will recall the distress caused by some aspects of the earlier publication[1], but chiefly that an important task had been undertaken and had been brought to a fruitful conclusion. The contributors, represented by today's panel, faced forward confidently to embark on a Six Subject Survey which includes, in addition to the three volumes already cited, studies of English and French as foreign languages, and civic education. The papers you have just heard, especially the last, have addressed themselves to the data on all six subjects and their analysis.

One may well begin by remarking that the three current volumes contain that wealth of technical detail that will please members of this professional audience of specialists in educational measurement and evaluation. Whatever may be said today, the data are there to be explored, analyzed and interpreted by those of you who, like me, look for meaning in the "anatomy of difficulty of the individual test item" as the key to the significance of the broader evaluations and generalizations built upon these measures to illuminate the effects of different patterns of instruction, school organization and holding power that prevail in the several countries that cooperated in the studies.

*Discussion of papers at NCME symposium on "The International Educational Achievement Study: Methodological Issues and Selected Results - Chicago, April 18, 1974.

In his paper today, Dr. Thorndike has fulfilled the role of the true test specialist in describing not only those problems in instrument development that were met and resolved to his full satisfaction, but others that had to be endured under pressure of time in the operation of the complex task of communication. The efforts involving translation and re-translation back into the original language are to be commended. One may wonder at the effect of the fact that English, a western language, was always the language through which the materials originating in other countries had to be processed (and reprocessed). Could this have influenced the finding that the developing countries were so far behind the developed countries in reading comprehension? Dr. Thorndike has freely admitted that the favorable findings of the pilot study of 1962 involving translating and retranslating through Turkish were not replicated for the reading passages in the current Reading Comrpehension Test. Since two of the three developing countries are eastern in language form and culture (Iran and India), could the students in these countries have suffered unduly? His own remarks suggest the likelihood of some effects of this sort for Iran. It could be that their lower performance is no greater than that of Chile, also a developing country, but with a western language and a smaller deficiency. No model for testing this hypothesis on the current data suggests itself and onemust wonder at the value of a separate study of items processed through Chinese or some other language of one's choosing.

Dr. Thorndike also raises the possibility of within-country semantic problems, but more about that later.

Dr. Wolf in his paper cites the problem of obtaining satisfactory evidence from ten-year-olds. This calls to mind the problem we faced in the Ford ETV experiment in Atlanta when we had to secure evidence from third graders on what they had learned from a TV science program. We did not use open-ended questions, but prepared 3-option multiple choice questions which the teachers read to the students while they had copy before them. The teachers read each question aloud twice, telling the students to listen for meaning the first time, then mark answers the second time round. Teachers learned to pause at the end of the stem or

question and at the end of each option. Children who could not read silently with comprehension knew to proceed from left to right and to move down whenever an option was completed. Despite the fact that we had not screened out many mentally retarded children that early in school, we found practically all papers were scorable, there were few dubious responses. Could taped oral instructions be used for such oral testing and the coding of free responses thereby be eliminated at the 10-year-old level?

The procedures of linear regression analysis are well described in Dr. Wolf's paper and in the separate volumes. One could hardly improve on the ordering of variables and blocks, or on the "screening" and "compositing" of the mass of independent variables available. The stepwise regression procedure seems also to contain its own corrective since, once a variable has been admitted to the total predictor set, it may fend for itself and not be limited by the sequential process. However, if the block procedure involves predicting only criterion residuals, it would only be necessary to run one ortwo sets of data through in reverse order of blocks to test the effect of order.

The use of the language of handicapping here and in the printed volumes may be justified, but one wonders why it would not be equally clear to speak of exceeding or falling short of expectations and predictions. Until landlocked countries like Switzerland and Bolivia are included in the studies, we may hope that yachting handicaps will be understood by most interpreters of results; but is this necessary?

Dr. Postlethwaite has packed into his paper and handout much data scattered over a larger expanse in the separate volumes. To take up his points in order, we may first underline his emphasis on the extent to which Reading Comprehension as measured contributes to the multiple regression prediction of all else. Stated in several different ways, Reading Comprehension is the primary predictor of Science and Literature test scores in the several countries and contributes almost as much to the prediction of Science test scores as to scores for Literary Comprehension and Interpretation. More on this later, but a high premium on verbal comprehension on the tests must be suspected.

His comment on the effect of school specialization on prediction for 18 year-olds calls to mind my brief venture into international education ten years ago. I was puzzled by the sharp bifurcation of education in Pakistan from grade 7 up into separate humanities and math-science curricula. My puzzlement was matched by that of my Pakistani counterpart who found great difficulty in comprehending the amount of general education going on in Athens High School when I took him to visit the principal. Statistically, this translates into negative weights in predicting Science scores from Literature scores, and vice versa, in countries where this early specialization occurs!

Dr. Postlethwaite next summarizes the regression analyses within countries in the handout you have before you. His speculation that learning conditions have a greater proportional influence in between-student analyses on Science, French and English than on Reading Comprehension, Literature and Civics because the former are more school oriented, may be broadened to include the fact that the former subjects are generally electives or subjects of specialization while the latter are among what George Stoddard[2] years ago classified as the "cultural imperatives." If mathematics had been included in this set of variables, one may confidently predict it would fall with science and the languages. From a study I reported in these meetings in 1961 and from other data available to me at that time, I can report that achievement test scores in elementary mathematics were more school-oriented, hence less variable, than scores in reading comprehension for the same school children.

The post script on parent and teacher effects and the tangential reference to Jencks leaves room for debate. That good teachers (and good parents) can be determined only after the fact by their effects is to disregard the accumulating evidence that variations on a "mastery learning" approach are powerful influences. I can testify from my own experience in teaching statistics to graduate students in education, leaving aside all

that Bloom, Lock and others have reported. I find Bennett Underwood's dictum that "a good teacher is one who gets his students to spend twice as much time on his subject" is amply true and permits one to observe such effects concurrently if not in advance by reputation.

It seems worth taking time to enlarge on a point Postlethwaite had only time to give a sentence to in his paper, namely: "It is also interesting to note that the between-school variance as a proportion of the between-student variance differed considerably. In general, in developing countries the proportion was highest." Here is a concept worth exploring and explaining. Is this again a function of specialization, this time by schools? It remains for such analyses to be reported and analyzed subject by subject. Only in the volume on Literature Education is this done explicitly and there the consistently highest proportions are shown for Italy, the lowest for New Zealand. A short summary volume, such as promised in Wolf's paper, would help to guide the reader to such generalizations across subjects as well as countries.

To turn to the three volumes in order, Volume I on Science Education includes findings that (1) confirm earlier evidence that boys excel girls in all the major sciences and "win going away", that is to say the differences increase with age, (2) show that retentivity cuts two ways, as it did in mathematics: high selectivity cuts failures and per capita costs, low selectivity maximizes success and diffusion of competence in the society, (3) present special problems of interpretation: as more able and more affluent students remain in higher age groups, the relative contribution of schooling increases because of decreased variance in environmental background factors, (4) are inconclusive in that unexplained variance in final achievement may be attributable to effects of schooling factors imperfectly measured. A significant methodological suggestion is that regression analysis be turned to a kind of case study use: that is, let schools be identified as exceeding markedly or falling far short of expectations based on regression, then let those schools identified be studied for sources of excellence or deficiency.

Volume II on Literature Education pursues a pluralistic concept of excellence, or even of truth. Careful treatment is given to consistent departures by students of a given country from answers deemed correct for the questions of comprehension and interpretation. Such consistency is accorded the respect due an autonomous culture. Beyond this it is admitted that the distinction between comprehension and interpretation items does not hold up in either the Literature Education Test or the Reading Comprehension Test. A specially conceived test of response preference to literature shows low reliability across nations and across readings, by the same token showing different dominant national response preferences and differences in response preference triggered by different reading selections.

Volume III on Reading Comprehension holds a special claim on our interest. We have come to recognize reading comprehension as a truly basic skill predictive of every kind of academic achievement at every level of education. But one may say this test succeeds too well. It predicts Science test outcomes almost as well as Literature scores. Why? Yes, why?

Over the years we have heard many criticisms of tests. Our own profession has been self-critical. We respect the judgments of our peers assembled in the Mental Measurement Yearbooks. We face our peers and try to do better. In the 1930's we took steps to overturn the excessive emphasis on memory for factual detail, epitomized in Stephen Leacock's quip that the Ph.D. meant you had been "examined for the last time and pronounced completely full." Chiefly under Ralph Tyler's leadership, we taught ourselves to measure "higher mental processes" by multiple-choice items. We developed best-answer items to go along with right-answer items. We conceived negatively worded items to test for understanding of multiple causation or multiple acceptable solutions or to test for least acceptable solutions. We learned the truly remarkable flexibility of this item-type. Life situations were seen to present multiple choices for decision.

Multiple-choice items require plausible "distractors," n-1 plausible distractors for each item with n options. Often we went the next step and said 5-choice items were preferable to 4-choice items because they reduced the guessing factor, or the probability of answering correctly from partial knowledge. Items in science and social studies tests became reading comprehension items. On, no! , merely required the careful reading the examinee should do in reading a textbook. But what about those "plausible distractors," all four of them? One of our public detractors of a decade ago used to make fun of multiple-choice items by proposing

"Paul Revere made his famous ride on a (an)

(1) motorcycle
(2) airplane
(3) automobile
(4) sled
(5) horse"

But we knew he was wrong, we didn't make items like that. Rather, we would propose

"A chemist working for a toothpaste company wishes to prepare 250 $cm.^3$ of 0.010 molar aqueous solution of stannous fluoride, SnF_2. Fortunately for him, SnF_2 is soluble in water. One mole of SnF_2 weighs 156.7 g. Equipment available includes a 250 $cm.^3$ volumetric flask; a 10 cm^3 pipette, a 0.01 g. sensitivity balance, and a 400 cm^3 beaker. Once the proper amount of SnF_2 has been weighed, which one of the following procedures would be best?

A. Place the SnF_2 in the beaker and add exactly 250 cm^3 of water from volumetric flask.

B. Place the SnF_2 in the beaker and add exactly 250 $cm.^3$ of water from the pipette in 10 $cm.^3$ portions.

C. Place the SnF_2 in the volumetric flask, dissolve it in less than 250 cm^3 of water, and then dilute to the 250 cm^3 mark.

D. Using the beaker and balance, weigh out exactly 250 g. of water and add the SnF_2 to it.

E. Dissolve the SnF_2 in more than 250 $cm.^3$ of water in the beaker, mix thoroughly, and then fill the volumetric flask to the line with the solution."

By this time the examinee is probably "mixed thoroughly." Need I say this is Item 29 on Test 10A for Population IV (Science Education in Nineteen Countries, p. 376).

But wait. What of my own examinations and yours? What have we done? Have we not presented items in each of several fields with an acceptable answer and 4 plausible distractors? Have we not presented difficult tasks for even the sharp reader? If the distractors are not highly plausible presentations of misconceptions, often only slightly incorrect, the examinee, we say, may answer by a process of elimination, ruling out options as wrong as those on the Paul Revere item.

As one step away from this hazardous practice, might we not abandon our psychometric predilection for a fifth option to reduce the verbal overload it introduces into an already trying experience? Item analysis will not ordinarily help us detect this effect although it will permit us to select the least efficient distractor in past use or experience with the item.

But consider further the problem we have created. Look at the typical reading comprehension exercise: a passage followed by several multiple-choice questions on the central thought, significant detail, application to another situation, generalization or inference, comparison with background knowledge, etc. What must the reader do? He/she must read and comprehend the passage, but to show that he/she "comprehends" must read and comprehend a whole set of questions or item-stems and, in each instance, four or five variously complex and fairly acceptable options, one of which can generally be considered best by a panel of highly qualified substantive specialists and verified by item analysis. We have compounded our demands in the name of valid assessment!

Who suffers? The reader for whom a test is given in his second language is one. Not only the foreign student in our midst, but in some developing countries at the higher grade levels in his own country because of the paucity of development of specialized terminology in his own language. I wonder about the IEA Population IV examinations

in India? Were they in Hindi or English? Ten years ago they would have been in English at high school in Pakistan and, I suspect, in India. I will never forget attending the dedication of a new eye-hospital in Allahabad, India, at which then Prime Minister Shastri was making the dedication speech. In the midst of considerable Hindi that I could not understand, I suddenly heard him say "medical aid and facilities" and, shortly later, "medical treatment." My American missionary friend explained that terms for these concepts had not been developed in Hindi because instruction at the higher levels was conducted in English and based on English instructional materials. Even those highly fluent in a second language are often so hampered as to require more time than allowed. I hark back to my oft-repeated statement[3] that our best cognitive measures developed in this century were the original USAFI Tests of General Educational Development, given veterans as they were being mustered out at the end of World War II. Uncle Sam generously allowed unlimited time. And when I followed Oscar Buros' injunction that I as a reviewer take the tests myself, I made the same high showing on Test III - Science Materials as I did on other parts in which I was more proficient by doing what I had done in college to make high grades in physics and chemistry: I spent twice as much time on those subjects as on any others.

So my second proposal is to remove the time limits from our tests so far as feasible. I have a standing proposal that the GRE Aptitude Test be offered so that those who wish might spend the entire morning on the Verbal Section and the afternoon on the Quantitative Section. If renormed under these more generous time limits, actually double the present limits, foreign students could do themselves justice in ways they cannot at present. Others from native backgrounds in which pressure of time is not ordinarily an accepted life style, including not only those from minority backgrounds but older students (35 and up), would generally opt for this choice.

A still more fundamental criticism remains. It is that the requirement of plausible distractors at a high level puts an excessive verbal overload on the whole examination process. Much of adult

professional life is conducted on a level of discourse that is serious and high, but not excruciatingly demanding of precise verbal comprehension. In such professional activity there is a premium on mastering and digesting great amounts of new material and concepts in this age of "Future Shock." To do this, a decent regard for precise meaning needs to be coupled with efficient learning. To persons of this habit our highly plausible distractors may contain so much of the truth, but not the whole truth and nothing but the truth, that they are distraught. In life, extensive reading and private conversations would often bring out the fine distinctions we cram into our multiple-choice answers. Faced with demanding decisions in the regular flow of related activities, these professionals make the necessary effort to be sound and successful at required intervals, but not in grasshopper-minded fashion from one fragmentary discrete item or problem or passage to another.

A footnote may be inserted here. True-false items, with all of their faults, are mercifully brief by comparison and stand on their own feet independent of the degrees of fine distinction presented by our multiple-choice format.

What can be suggested? Turn to our technology for computer-assisted instruction, videotape feedback, even audiotapes of conversations or broadcasts. Is it not time that we mounted a full-scale basic research effort designed to get at fundamentally valid appraisal of understandings and mastery of behavioral objectives rather than gradually (grudgingly?) yielding to pressure groups who would throw out the baby (objective measurement) with the bathwater of unnecessarily biased determinations and interpretations of performance? It is true that it is difficult to demonstrate empirically more reliable and more valid technological devices (that cost more) on a short-range cost-benefit basis.

But let us turn back the pages of measurement 50 years when I was taught that highly speeded tests of cognitive ability were to be preferred to power tests because the correlation between speeded and unspeeded was .90 or so. Today we use what we intend to be power tests

to the exclusion of speeded tests except where speed of response is a field requirement. We have only to turn to studies of Schrader[4] and his associates around 1950 in which he compared the predictive validity of the ACE Psychological Examination for fresman college achievement with that of the Selective Service College Qualification Test of that time. For 29 independent comparisons the latter unspeeded test reliably outran the well-constructed but definitely speeded ACE Psychological, the average correlations being .55 and .43, respectively.

May we then meet the challenge of critics of our honestly conceived, but admittedly imperfect present measures by demonstrating the validity of verbally uncluttered, completely unspeeded, pictorially and otherwise technologically facilitated direct measurement of educational outcomes. International studies seem to be a place to start because of the special demands they place on our current instrumentation.

1. Wilson, James W. and Peaker, Gilbert F. "Special Issue on the International Study of Achievement in Mathematics." Journal for Research in Mathematics Education, Vol. 2, No. 2, March 1971. pp. 69-171.

2. Stoddard, George D. The Dual Progress Plan New York: Harper and Row, 1961, 222 pp.

3. Findley, Warren G. Review of the "USAFI Tests of General Educational Development" Third Mental Measurements Yearbook, pp. 42-46. New Brunswick, N.J.: Rutgers University Press, 1949.

4. Educational Testing Service "Statistical Studies of Selective Service Testing" SR 55-30. Princeton, N.J.: Educational Testing Service 1955.