DOCUMENT RESUME

ED 111 881 TM 004 849

AUTHOR Thorndike, Robert L.
TITLE Methodological Problems in Developing Instruments for Cross-National Studies.
PUB DATE [Apr 74]
NOTE 13p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, Illinois, April 1975); Not available in hard copy due to marginal legibility of original document

EDRS PRICE MF-$0.76 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS Communication (Thought Transfer); *Comparative Education; *Cross Cultural Studies; Cultural Differences; Curriculum; *Test Construction; *Testing Problems
IDENTIFIERS International Evaluation Educational Achievement

ABSTRACT

In developing tests for the International Association for the Evaluation of Educational Achievement (IEA) survey, methodological problems arose in three areas: curriculum, communication, and culture. Efforts to identify the core of common objectives and the penumbra of distinctive, sometimes partly shared but sometimes unique, goals operated through a system of national and international committees. Each country was given the responsibility of assembling a national committee, each having the task of preparing a national blueprint of content and process objectives that would be appropriate at the specified age or grade levels in that country. Through interaction with national and international committees, items were selected, edited, and assembled into preliminary forms for try-out. Communication was a problem in maintaining the flow of information, materials, and actions out to the participating countries and back to the central coordinating office. In a more specific sense, communication was a problem in the domain of language and translation. Problems involved in the area of culture were semantic and in picking a set of quantitative alternatives giving good differentiation between countries. (Author/RC)

Methodological Problems in Developing Instruments for Cross-National Studies

Robert L. Thorndike

Teachers College, Columbia University

If I had to produce a capsule summary of our methodological problems in developing the IEA instruments, I would say, "Curriculum, Communication and Culture." Let me expand on this to provide clarification and substance.

Whenever a test is to be given to evaluate educational achievement, it is important that the test tasks match the learning outcomes that are set as objectives of the instructional program that is being evaluated. This is the familiar notion of content validity drummed into every student in his introductory testing course. It gets fancied up with lists of behavioral objectives and criterion references, but it is still the ancient maxim of "test what you teach."

Achieving a precise match between instructional objectives and test tasks presents problems even within a country if there is a degree of decentralization and diversity--as there emphatically is in the U.S.A. What is the main theme in one social studies program, for example, may be perceived as peripheral or even irrelevant in another. But the diversity seems likely to be compounded if one deals with 10 or 15 or 20 countries. How shall one deal with that diversity?

The problem has two sides: (1) How shall one determine the dimensions of the diversity? (2) Having identified the community and

the diversity of objectives in different countries, how is one to deal with what one finds?

In the IEA studies, our efforts to identify the core of common objectives and the penumbra of distinctive, sometimes partly shared but sometimes unique goals operated through a system of national and international committees. Each participating country was given the responsibility of assembling a national committee, presumably well versed in the curriculum of math or science or reading instruction in that country. Each national committee had the task of preparing a national blueprint of content and process objectives that would be appropriate at the specified age or grade levels in that country. The national outlines were to be fed in to a central international subject matter committee that had the responsibility of collating them, identifying areas of agreement and areas of divergence, and then proposing a composite international blueprint. This was then returned to the national committees for review, criticism, and suggestions for modification. With varying amounts of interaction back and forth, the content by process blueprint was stabilized in a final form.

The same type of reciprocal interaction was to take place in the preparation of test exercises. That is, the national committees were invited to submit possible exercises to an item pool, and these were reviewed by the central international committee. A selection of possible items was made, and these were sent back to the national centers for review and comment. In the light of such comments as were received, items were selected, edited and assembled into preliminary forms for

try-out.

This, at least, is how things operated in theory. But if you know anything about humankind, you know that national centers varied widely in the promptness and in the meticulousness with which they responded to requests for materials or for reactions to materials. Thus, inputs from national centers tended to be spotty, with some having much more influence than others on the final product, and a disproportionate share of the determination of what appeared in the final tests fell upon the central international subject committees. The logistical problems of maintaining an effectively functioning world-wide communication network for a project of this sort are very severe indeed.

One strategy would say: Build a separate test for each country, to match that country's objectives. This is a conceivable strategy if one thinks of countries solely as opportunities to replicate in different settings some strictly intra-national types of analysis. If, for example, one wanted to study in a number of countries relationships of sex of teacher and sex of student to mathematics achievement (assuming that this were a problem worth studying), it would not seem important to use the same identical math test in each country. Different tests, each tailored to the objectives of the specific country, would seem to provide legitimate evidence on a problem such as this. It is possible that the specific content of the test would interact with sex of teacher and student, but it seems unlikely. However, if the enterprise is concerned in part with comparing the levels of achievement reached in different countries, there would seem to be no way to do

this except through a common set of test tasks. What, then, should be the specifications for these tasks? At the two extremes, they might be either (1) limited to tasks that correspond to objectives espoused by all countries or (2) extended to include all objectives espoused by any country. An intermediate position would be to plan to assess objectives agreed to by several but not all participating countries.

No one of these choices is ideal. Limiting the assessment to universal objectives is likely to produce an excessively narrow test, and one that is least adequate for the system with the most inclusive curriculum. Including the complete range of objectives implies testing students in some countries on many topics on which they have had no instruction. An intermediate stage represents a compromise between these two ills, but not the elimination of either of them. Incidentally, I believe that this compromise solution is the one that IEA adopted in most of the cases. It is also my impression that the situation was not quite as desparate as I have made it sound, since in large part the content and objectives in mathematics or science or reading were common across countries. A further adaptation to the differences that clearly did exist in balance and emphasis was to provide part scores and item statistics, so that a country's achievement could be compared with the others not merely on total mathematics score, for example, but on arithmetic, algebra and geometry, or on computational skills vs. problem solving. National profile patterns were in some ways more instructive than national standing on the "educational Olympics." One final adaptation was to get in each country estimates of how com-

monly students had been taught the content covered by each item, and to use this measure of "opportunity to learn" as one independent variable in a number of analyses.

My second key term was "communication." This was a problem in two quite different senses. One I have already alluded to. This was the logistic problem of maintaining the flow of information, materials, and actions out to the participating countries and back to the central coordinating office of the project. It is hard enough to try to keep a single national survey, directed out of a single national head-quarters, operating smoothly and on schedule. Adding an additional layer of coordination on top of this, with additional flow of information and materials back and forth across oceans and continents at each step in the way makes maintenance of an established schedule of operations almost impossible of fulfillment. We learned of the difficulties as we went along--of floods in Hungary and epidemics in Aberdeen, or mark-sense cards lost in transit or swallowed up by Customs, of well-intentioned national centers that never did get the try-out booklets administered. We came to realize the absolutely vital importance of a strong international office, with a compulsive administrator to monitor the flow of information and material.

In the most recent cycle of studies, we adopted the strategy of having in each country a nearly full-time National Technical Officer, who provided the responsible dynamic within the country to meet commitments and deadlines. We were impressed with the necessity of spelling

out all procedures and schedules in operating manuals that were infinitely detailed. We came to rely upon intensive week-long briefing sessions of the National Technical Officers at which all procedures were reviewed and even the most minor details worked out. But even so, participation in planning and review were spotty, and we still had one or two instances in which operational slippage occurred--such an unhappy event as an item being mis-keyed, or a country testing fifth graders instead of 10-year-olds.

The other sense in which "communication" was a problem was more specifically in the domain of language. In the survey of achievement in science, in which we had the greatest number of participating countries, it was necessary to translate all materials into 14 different languages ranging from Finnish to Japanese. The translation was required not only for the tests but also for questionnaires for students, teachers and school officials, and in addition all the manuals and procedural guides that directed the work of the coordinator in a school system and the test administrators who actually carried out the testing. It was a horrendous task!

At this point the question arises: How adequate was the translation? Did a given test exercise present the same task after translation into each of the languages? Did the background questionnaires present in all essential respects the same questions to children or teachers in each country? How does one know? I should note in passing that English was the common language through which everything passed on its way to the other languages. That is, if the Finnish National Center contributed a biology item, it was translated from Finnish into English before being translated into Italian, Japanese, Hindi,

Thai and all the others.

It is perhaps for the reading tests that one becomes most concerned with problems of translation, since in these tests language appears to be of the essence. What evidence can one present that the test task has not been subtly or even grossly distorted by the process of translation?

Our original hope had been to get an immediate and independent back-translation of all of the passages and items, and to use this to police any distortions that might seem to have crept in. Alas, neither time nor resources of translators were available to make this possible. We do have back translations of selected passages, together with their items, but these were received after the fact, and could not be used to make any modifications of the tests.

Two lines of evidence from prior studies had led us to believe that translation problems might not be too serious. One has to do with the consistency of relative item difficulty from one language to another. We had included a little reading test in our initial pilot study reported in 1962. In this study the correlation from language to language of item difficulties, expressed as percent getting the item right, was 0.90 and this high correlation seemed to suggest that each item maintained its character with little change under translation. A second line of evidence comes from a Teachers College doctoral dissertation studying the possibility of using the combination of a reading test in English and one in the native language (in this case Turkish) as a basis for appraising both scholastic aptitude and degree of mas-

tery of English of foreign students who might come for college studies in the U.S.A. The cross-language difficulty indices din't correlate as well in this case--about 0.70--but a back translation was produced. In this study no significant differences in difficulty were found in mean scores on the original and the re-translated versions of the tests when given to high school students in the U.S.A. For one form, the correlation of item difficulties between original and re-translated form (corrected for the unreliability of the indices) was 0.95, while for the other form it was 0.77. Thus, the items and tests did not seem to have been too badly distorted by translation into Turkish and back again.

So we went ahead and translated the materials not only for the tests of Mathematics, Science and Civic Education, but also the passages used to measure reading comprehension and literary comprehension and appreciation. It is only for the Reading Comprehension Test that I have had a chance to examine the consistency of item statistics from language to language. Alas, the correlations are not as high as those that we found in our pilot study. The average cross-language correlations of item difficulty were approximately 0.75 for 10-year-olds, 0.70 for 14-year-olds and 0.65 at the end of secondary school. For item discrimination indices the corresponding correlations were about 0.60, 0.40 and 0.45.

The results suggest that maintaining comparability under translation becomes a progressively more serious problem as the material to be translated becomes more difficult. This is perhaps not surprising.

It may arise from either or both of two influences. On the one hand, simple ideas and simple items may have more exact counterparts in other languages. On the other, simple materials place less of a strain upon the cognitive and liguistic skills of the translators. Thus, the most difficult passages were ones that had in the past been used as part of an admissions test for doctoral students at Teachers College. It would not be surprising if even a very capable Iranian educator, for example, whose native language was not English, had difficulty in rendering precisely in Persian a passage on the philosophy of science or the determination of gross national product. I have a sneaking suspicion that reading a back-translation for a few of the most difficult passages, if they had been prepared, would have been a somewhat gruesome experience.

We attempted to carry out a scrutiny of those items in which certain countries showed sharply deviating responses--deviating especially on the error choices that they selected. Our effort was to understand why the discrepancies arose. We asked the National Technical Officer in each country to give a rationale for each of the peculiarities of response in his country. We asked him to try to judge whether the peculiarity arose from some idiosyncracy of the national language or from some idiosyncracy of the national culture. But the effort wasn't very productive. The judges expressed very great difficulty in making the judgments, and the rationalizations that they offered were singularly unconvincing. The only <u>really</u> convincing explanation arose in one or two instances in which they had reversed the order of the op-

tions, or made an error in the scoring key.

Mention of culture brings us to the third potential problem in preparing instruments for use in various countries. Are the tests, and especially the questionnaires, suited to the culture of each of the countries involved? For example, one reading passage concerned Ernenek, an Eskimo boy, who lived in a snow igloo on the top of the world and "iced" the runners of his sledge to make them slide better on the ice and snow. How does a passage of this type perform in Finland and Sweden on the one hand, which were the most nearly arctic of our countries, and the Netherlands and Chile on the other, where it is unlikely that anything remotely resembling an Eskimo or a sledge has ever been seen? It is comforting to find that Finland and Sweden do relatively no better on this passage than others, and the Netherlands and Chile relatively no worse. I have not made a systematic check passage by passage, and this should probably be done to see whether national variations on specific items are peculiar to the item, or reflect something more general about the passage as a whole.

On the questionnaires, some problems arose relating to the wording of the questions. However, the major difficulties centered on the response options. In order to keep the data reduction within manageable limits, every effort was made to pre-code the options on the questionnaire completed by the students, teachers and a school administrator. A given response option needed to be uniform across all countries if the data were to be reduced to alphabetic or numerical codes, consol-

idated within countries and compared across countries. But in preparing these codes two types of problems were encountered. These will be illustrated by some fairly representative examples.

The first type of problem was semantic. Consider the question: "Which of the following best characterizes the community served by this school?" The alternatives in the English version are various combinations of "urban," "suburban," and "rural." It seems likely that "urban" and "rural" will have fairly uniform meaning, but are "suburbs" as we think of them a meaningfull concept in all cultures? Or again, in a question about the amount of training in physics that a science teacher has had, how does "between 2 and 4 semesters" convert into the training programs in England, or Hungary, or Iran, to say nothing of a U.S. university on the quarter system.

The second type of problem relates to picking a set of quantitative alternatives that gives good differentiation between countries. This can be illustrated by the question: "How many books are there in your home?" Response categories ranged from a low of "None" to a high of "More than 50." These options worked well in countries such as Chile and India, but in Sweden about 80 percent of the respondents marked the highest category, and there was, as a result, very little spread across the group of Swedish respondents.

Of course, all the questionnaires encountered the full range of problems that plague questionnaire and survey studies within a country. Options appeared not to be applicable in individual cases. Many schools appeared to have only impressionistic data on expenditures

within their school. One may question the accuracy of student responses to questions about parental occupation and education, though some preliminary studies indicated that pretty good correspondence was obtained between student and parent reports. These internal problems become accentuated by the difficulties in maintaining equivalence of meaning across languages and cultures. Thus, relationships (or the lack of them) between family and school factors and the dependent variables of school achievement need to be scrutinized critically by the researcher in the country involved to examine the possibility that unexpected results may represent some deficiency in the instrument, rather than a genuine peculiarity of the particular educational system.

In my presentation I have focussed on the methodolgoical problems. Obviously, we have felt that we have arrived at tolerable solutions to these problems, though far from ideal ones, because we did proceed with the study. But reviewers of the findings must remember that this is a large scale survey type of study, with all the limitations in types of data and integrity of the results that this implies, and that in a cross-national study these limitations are doubled in spades.