Needed: Instruments as Good as Our Eyes*

Originally published as Paper #7, Occasional Paper Series, July, 1976

* This paper was originally published in the *Journal of Career Education*, Winter 1976, 2(3), 56-66.

Henry M. Brickell

> "All right. Pick up your yellow pads; pack your suitcase; get on the plane to Ohio. And don't come back. Not until you have an instrument that can measure as well as your eyes, ears, noses, and throats."

I was tired. And frustrated. Why couldn't they measure what they could see? If career education in Ohio was as vivid as they said—if teaching was as career-flavored–if students learning was as apparent–why couldn't our evaluation staff measure it? Why couldn't we have hard evidence to match soft evidence: test scores to match classroom observations?

Mercury between the Fingers

We had been at it in Ohio for three solid years already. Teams of observers could see it, or so they said. Not only career teaching. Career learning. They could see it. But they couldn't pick up the evidence and bring it back to New York. Palpable as mercury but just as elusive: Independent observers could agree the silvery globules were there yet not one of them could pick one up and hand it over to another observer, or to us, or to Ohio officials, or to Ken Hoyt, or to America.

Thousands of students taught by hundreds of teachers in dozens of schools; interviewed, watched, and tested. All kinds of teachers, young and old, good and bad; all kinds of students, bright and dull, black and white; all kinds of schools, urban and rural, rich and poor. Learning, learning everywhere but not a statistical difference.

Thorndike echoed across fifty years: "...if it exists, it exists in some quantity and can be measured..." or something like that. Easy enough for you to say, E. L. If you were here, I'd give you a yellow pad and put you on the plane to Ohio. See what you could do. It exists, all right. Go measure it!

Four Years Ago: Observe It

Four years ago, we hadn't been sure. Back then, Ohio had issued a career development curriculum guide in three volumes, very impressive: Career Motivation for K-6, Career Orientation for 7-8, Career Exploration for 9-10. Our mission as external evaluators: try to find it in the classroom.

Out went our teams, hardheaded people, skillful observers, familiar with the ways of career education, out across the old Northwest Territory, following the State guides. They coursed down the Ohio to Cincinnati, up the Miami to Dayton, up the Scioto and along Paint Creek to Paint Valley, up the Muskingum to East Muskingum, over land north to Akron and still further along the Cuyahoga to Cleveland. The team sat with project staffs, looked at materials, talked with teachers, visited classrooms, watched students. In time, the word worked its way back East.

"We have found it: project staffs explaining careers, teachers teaching careers, even students learning careers–especially in K-3, a bit less in 4-6, still less in 7-8, and least in 9-10. It cannot be found everywhere and it differs according to habitat and natural surroundings, but its range is statewide."

At year's end, we told the State officials in Ohio. They were pleased with what their $25 per pupil had accomplished in the K-10 pyramids of project schools–K-6 feeding into 7-8 and on into 9-10. We were happy; they were happy.

Then they said the words that have destroyed many happy marriages between friendly observers and friendly program people: "Test it."

Three Years Ago: Test It

Easy. We knew exactly how to do that. Take the State Curriculum guides, thumb to the objectives, write multiple-choice test items to match, show them to State and local officials as a face validity check, pilot test the items to make sure the difficulty level is right, and compile final tests for grades 3, 6, 8, and 10–the natural terminal points of major program segments. We did all of that.

Then we administered those tests to 6,000 students–students in the program and students not in the program–in four representative cities, cities our teams had visited. And then we analyzed the results. What did we find?

No significant difference. None. Not in cognitive learning. Not in affective learning. Not in item clusters. Rarely in individual items.

Mercury

Of course, we were not alone. Soon the nation would have spent $10 billion in ESEA Title I since 1965. $10 billion while 10,000 teachers and specialists and observers insisted that something good was happening: teachers were teaching, paraprofessionals were tutoring, students were learning. But Jim Coleman said "No significant difference." And Christopher Jencks seemed to agree. Very familiar. Too familiar.

The Ohio officials were not pleased. That year's finding canceled the previous year's finding. Then you saw it, no you don't. Magic. Magic that makes the client's money disappear, leaving nothing. No findings. Not even an explanation.

How could this be? How could our independent observers have gone out, seen nothing, and called it something? They were too good to be wrong; they had observed the way they were supposed to. Still, our tests were too good to be wrong; we had made them the way we were supposed to.

What were we to believe: The evidence of our eyes to the evidence of our instruments. And how, as supposedly-skilled evaluators, could we explain the conflict?

Four explanations occurred to us. One assumed that our instruments were right; one assumed our eyes were right; two assumed both were right. Here were the possibilities:

Interpretation 1: No career education is taking place. Our observers are blind–or hallucinating. Our tests are right.

Interpretation 2: Non-career curricula teach career content. Program teachers are teaching it and program students are learning it, but so are non-program teachers and students. Our observers and our tests are both right.

Interpretation 3: Non-school sources teach career content. Career information is so important that the home does not leave it to the school. Family, friends, and TV teach the content to non-program students. Our observers and our tests are both right.

Interpretation 4: Program students cannot fully demonstrate their learning on our tests. Career education is taking place. Our observers are right; our tests are wrong.

We knew that Interpretation 2 and 3 were right. Non-program students scored well on the tests, tests that Ohio local and State officials said fit their curriculum. Somebody in school or out of school was teaching career concepts to non-program students.

But what about Interpretation 1 versus 4? We decided to go back to Ohio, look at those classrooms, double check our impression not only that teachers were teaching but that students were learning.

Two Years Ago: Observe It Again

Out went the observer team again, down the Ohio, up the Miami, along the Cuyahoga. This time they went to seven sites–a good cross-section. We waited for their report.

When it came, it sounded stunningly familiar, completely reassuring, and deeply disturbing.

"We have found it again: project staff explaining careers, teachers teaching careers, even students learning careers–especially in K-3, and bit less in 4-6, still less in 7-8, least in 9-10. It cannot be found everywhere and it differs according to habitat and natural surroundings, but its range is statewide."

Thanks a lot! I guess.

So they were not blind. Or at least they hallucinated consistently. But how could we prove that they were not seeing mirages, illusions rising from the hot enthusiasm of project staff, teachers, and students eager to show the Wise Men from the East what they had come so far to see–and eager to keep their $25 per pupil? That is where last year's story began.

One Year Ago: Test It Again

We were back in New York, reviewing the evidence. We had seen it once, seen it

twice, but we couldn't make a test to show it at all. Not even with tests our own staff had built to match the Ohio curriculum–tests the local and State officials had endorsed.

Let's look again at Interpretation 4 and go over it once more, this time slowly. Let's see: Program students are learning something but they can't show it on our tests. Hmmm...okay...adopt that interpretation...try to prove it.

"All right, staff. Pick up your yellow pads; pack your suitcases; get on the plane to Ohio. And don't come back until you can measure what u can see. Pick up the mercury. Bring it home. Somehow."

"Try this. Leave those three Ohio curriculum guides in your offices here in New York. When you get on site, walk past the project staff, walk past the principal's office, slip past the teachers and sit in the backs of the classrooms."

"Watch the learning. Don't watch the teaching. If you ever see a student or a class learn something about careers–anything about careers–take your yellow pad and write a test item she could pass. Wait a minute. Now, this could be critical for us: make it an item no other student could pass unless she had been in that same class. Or in an equivalent career education class across the hall, across the town, across the county, or across the state. That's it: we want items not only for the rooms where they were written: we want items that will work statewide. We want items career students can get right but other students can't–items that show the special things only career students are learning."

"Write the item on the spot, if possible; at least write notes for it. Polish the items that night in your motel rooms."

"Ideally, we ought to administer those items to students the day we write them, or the next day. In fact, we shouldn't even bring items out of the classroom where they were written, if students can't pass them. We'll get items like that when you thought you saw learning, but didn't. However, we can't work that fast.

"So instead, as soon as the items are written, polished, edited, and printed, we'll administer them not only in that class but in other program classes in the same city–and to non-program classes as well."

Take along three big canvas bags: a green one for items that discriminate in favor of program classes, a red one for items that discriminate against program classes, and a grey one for items that don't discriminate. We'll give gold stars for green items."

"Now that is one way to write items. But only one. Let's think up others. How else can we get items that will let the career students show their stuff?"

The conversation went on and on and on. When it was over, we had thought of eight more. A total of nine ways to pick up mercury. Before the teams came home they would need them all.

Field Based Test Development

We were out to find whether tests developed in the field rather than if the office–tests to measure what teachers were teaching rather than what program designers planned for them to teach, tests to measure what students were learning rather than what program designers intended them to learn, tests that were classroom-sensitive rather than curriculum-sensitive–whether such tests could distinguish between program students and non-program students. That is, we were out to find whether we could measure what we could see. And the technique would be simplicity itself. We

would make the instruments where we have seen the learning– in the classrooms. We would build the tests to match the learning.

The item-writing teams spent the winter in the river valleys of Ohio, picking their way through the white snow and the brown slush, going from school to school and room to room, moving along the rows, picking an item here and an item there, filling their green bags and their red bags and their grey bags one item at a time.

It was slow work. A few of the nine methods worked; most didn't. Teachers weren't very good item writers; students weren't either; our writers weren't either. They strained hard to find career learning in grades 9-10. They strained harder to find affective learning–the much-hoped-for program outcome, according to State and local officials–at any grade. Maybe affective learning is as ephemeral to the eye as it is to an instrument.

They wrote over 1,000 usable items, leaving hundreds more in the wastebaskets of Ohio motels. Each item piloted with hundreds of program and non-program students in grade 3 or 6 or 8 or 10. Each was piloted with students in the city where it was written and in three other cities.

Three Bags Full

The writers brought three bags of tested items home to New York. We had to help them carry the grey bag, sagging with non-discriminating items. The green bag was much lighter. The red bag was, happily, the lightest of all.

We dumped out the contents of the grey bag and the green bag and compared them. The teams had to write 5 cognitive items to produce one discriminating in favor of the program and 9 affective items to produce one discriminating in its favor. Hard work, picking up mercury.

"Save all the items. Each one is equally valuable, equally diagnostic. We'll go over each one with the State officials in Ohio. We have to tell them–even if they won't be happy to hear it–where the program is strong, neutral, and weak–the green items, the grey, and the red." What was it Ohio State wore to the Rose Bowl this year–red and grey. We could have told them.

We took the greenest of the green items and compiled them into four tests–one each for grades 3, 6, 8, and 10–each with cognitive items and affective items.

We shipped them off to seven sites: somewhere we had written items; somewhere we had piloted them; somewhere we had neither written nor piloted them. We administered the tests to 12,000 students–some in the program and some not in the program, the two groups as comparable as we could make them under natural field conditions.

Significant Differences at Last

The new test results demonstrated the superiority of program students to non-program students on every test in every grade. They were ahead on the cognitive tests and on the affective tests; they were ahead at grade 3 and 6 and 8 and 10. Statistically significant differences at the .01 level in statewide results on every test in every grade (except grade 10 affective learning, where both our observers and our item writers had found the atmosphere thin).

Our eyes had been right. Our original instruments had been wrong. And we knew why.

But there was more. The results showed the program was working best in K-3, a bit less in 4-6, still less in 7-8, and less in 9-10—but it clearly was working at every grade. Exactly what our observers had said!

And more. The program was working in most grades in most cities but not in all and it varied in scope and success. What had our observers said? "It cannot be found everywhere and it differs according to habitat and natural surroundings, but its range is statewide."

And more. The test items written in one program classroom worked in many other program classrooms–across the hall, across the town, across the county, across the state. So there was much commonality statewide along with the diversity. It means teachers were teaching similar things, even though not necessarily the things in the State curriculum guides. It means that State officials had issued a common message, that local officials had heard it, that classroom teachers were listening to it.

Checking Validity: Before we scored the tests and analyzed the results, we asked the State and local officials in Ohio to rate each item in each test as measuring "important" or "neutral" or "unimportant" career knowledge and career attitudes in grades 3, 6, 8, and 10. Making their judgments independently, the officials rated an average of 87 percent of the test items as "important", 9 percent as "neutral", and 4 percent as "unimportant". So much for validity. That was the best we could do that year.

Checking Reliability: After we had scored the tests, we computed split-half reliability coefficients for each test using the Kuder-Richardson 20 formula. Affective test reliabilities ranged from .52 to .68; cognitive results good enough for individuals.

Using Spearman-Brown, we found that we would have to lengthen the 20-item affective test to 80 items to make them reliable as the 40-item cognitive tests. Not surprising. Feelings are more subject to variation from time to time than knowledge; thus our affective tests would have to be longer than our cognitive tests to be equally reliable.

Checking Item Difficulty: Both State and local officials in Ohio had wanted the test items to be made easy enough so that students could demonstrate clearly what they were learning. Many teachers echoed that concern. Our analysis showed that the final tests were indeed rather easy, with students averaging 60-70 percent of the items correct.

Checking Test Coverage: The Ohio career development program has seven divisions–world of work, economics, decision-making, etc. Our analysis showed that the greenest of the green cognitive and affective items came from all seven divisions, demonstrating that the program students were able to show their superiority across the entire program spectrum.

Convergence at Last

We could measure what we could see at last. The evidence of our eyes converged with the evidence of our instruments. And now it was easy to see why. Evaluators use their eyes to see what is there, whether it is intended or not. But they use their test instruments to measure what is intended, whether it is there or not. Like other evaluators, we had made field observations of the in-fact curriculum as it was being taught. And like other evaluators, we had originally written test items in the office to match the official program intentions.

We had achieved convergence by moving our item development out of the office and into the same locations as our observations. That suggested another route to convergence: train our observers to see as little as our tests. Train them to look only for what is intended and to ignore everything else that is accomplished. Just like our current tests.

Of course, it is important to know whether intentions were accomplished. But it is equally important to know what was accomplished, intended or not. Our Ohio curriculum-based tests measured intentions; our field-based tests measured accomplishments.

Indeed, the finished field-based tests themselves profiled the superior learning of the program students. The items themselves actually mirrored their learning. Say that again. What mirrored the learning—what? The items—the silvery test items spread on flat sheets. Exactly. The mercury at last. We had it. A mirror as good as our eyes.

We need such instruments. As evaluators, we need to be able to say to program directors and classroom teachers: "Yes, we can measure what you can see." Otherwise, we may look irrelevant or incompetent or dangerous.

Evaluators have been broadening their repertoire of instruments for years: curriculum-embedded tests, observer checklists, audiotape recorders, videotape recorders, unobtrusive measures, the critical incident technique, situational tests, peer ratings, projective tests, criterion-referenced tests, and on and on. Developing tests in field situations promises to enrich our repertoire still further.

The notion of field-based testing blends nicely with some recent and current thinking about alternatives to testing the declared objectives of a program.

Goal-Free Test Development

Unthinkable. Until you think about it. Michael Scriven suggests somewhere that the difference between main effects and side effects are in the eye of the beholder. Scriven says the program person calls what she intends main effects and what she does not intend side effects. The liberated goal-free evaluator, on the other hand, might call them "unifects." The goal-free evaluator's entire business is effects, not intents; out-comes, not goals; the target hit, not the target set.

But how could the evaluator ever develop a test? Well, she could look at the treatment rather than at the objectives. Since the treatment so often loses sight of the target anyway, she could ask "Where are you shooting?" rather than "Where are you aiming?" That is, she could start at the muzzle, and predict the trajectory, and go to the other end of the arc—rather than standing out there beside the official designated target, waiting for the holes to appear.

We are talking about developing test items by examining instructional processes rather than stated goals and objectives. And we are saying they might be a better source. Better in the sense that the evidence of the evaluator's instrument might more closely match the evidence of the evaluator's eyes. Better in the sense that they would report whatever is going on. Better in the sense that they would be firmer guides to the evaluator than the kind of "officials" pious overstated goals and objectives loaded like so much ballast into proposals to get them through the heavy seas of the review process, then dumped overboard.

Theory-Based Test Development

> Theory-based evaluation should be employed where an instructional program aims toward distant or intangible outcomes...
>
> Remoteness and in concreteness of objectives seem, in fact, to be particularly characteristic of the humanistic trend in education...
>
> The evaluation question becomes: Have the variables which theory indicates are crucial to the program actually been operationalized?...

This is what Carol Tylor Fitz-Gibbon and Lynn Lyons Morris said in the June, 1975 issue of Evaluation Comment (UCLA Center for the Study of Evaluation). Making a case for theory-based evaluation, they suggested that both formative and summative assessment should be guided by the theory on which the program is based.

> One cannot measure or observe or report on everything about a program; inevitably, one selects...
>
> The choice of variables to study need not remain a matter of opinion...
>
> When a theory-based evaluation is planned, the variables selected for study are those which a theory indicates are crucial...

Fitz-Gibbon and Morris talk mostly about process variables but they suggest that the same reasoning applies to outcome variables.

If they are right, how does one develop a test? Well, the theory underlining the program is—to put it over simply—a set of cause-and-effect statements about what processes will lead to what outcomes. Follows that a program based on that theory—irrespective of whether the program knows what theory it is using—is likely to produce certain outcomes whether it intends them or not. Presumably, those are the outcomes a skilled observer would detect with his eyes and ears. And those are the outcomes a skilled evaluator would create instruments to measure. Then, happily, the evidence of the observer's eyes and evaluator's instruments would converge.

"Don't bother to declare your objectives. We'll just check your program vehicle, decide which theoretical road it is designed to travel, draw a finish line across the end of the road, and wait there for you to arrive." Thus speaks the pure theory-based evaluator. The finish line is of course the criterion level and the ribbon stretched across the road above it is the instrument.

Field-Based Evaluator, Meet Goal-Free Evaluator and Theory-Based Evaluator

F.B. Evaluator and G. F. Evaluator and T. B. Evaluator can get along well together. None of the three is much interested in intended outcomes. All three predict the outcomes—F. B. from observing learning, G. F. from observing processes, T. B. from identifying theory. None of the three would design instruments to measure the intended objectives exclusively.

And all three would hopefully produce hard evidence that would coincide with soft evidence.

Protecting Career Education from False Conclusions

How good it might be for the evaluator to watch what career educators are doing, determine where that will take the students, and design tests to measure how far they get. Beyond that, evaluators might keep career education out of undeserved danger by telling project directors things like: "No. We won't test academic learning to see whether your one-day values-clarification workshop for teachers last summer raises students' scores on the Iowa achievement test next June. The treatment is too weak for that objective. In fact it isn't even trying. If you're going to clarify values all year, we're going to test values-clarification."

As evaluators, we ought to get our values clarified about that.