

## **On the Frontlines**

### **Evaluation Blues**

How accountability requirements hurt small, innovative programs the most

By Laura Silverstein & Erin J. Maher

Stanford Social Innovation Review  
Winter 2008

Copyright © 2007 by Leland Stanford Jr. University  
All Rights Reserved

# Evaluation Blues

*How accountability requirements hurt small, innovative programs the most*

by Laura Silverstein & Erin J. Maher

**T**he families living in the Vintage Park apartments in King County, Wash., have two things in common: their daily hardships and their passion for their children's success. Almost all are recent immigrants. Most live without enough food and clothing. Many speak only P'urhepecha, the language of an indigenous community in Michoacán, Mexico. All live in an apartment complex built in the 1940s as temporary housing for World War II veterans that now hosts mold and occasional gas leaks.

Isolated, overworked, and poor, these families are not sure how to help their children get ahead in the United States. Our organization, Burien, Wash.-based New Futures, tries to help them by offering after-school programs, assistance in meeting basic needs, and community-building activities. We strive to be

flexible so that we can meet the ever-changing needs of our diverse clients. We strive to be innovative so that we can use their strengths creatively. And we strive to be holistic by having all New Futures staff – teachers, social workers, and community developers – work together to integrate services for children, families, and the apartment community.

Although quantifying the outcomes of flexible, innovative, and holistic programs like ours is difficult, we have tracked our progress for a decade. But now we face mounting pressure to prove, with scientific precision, that our programs positively affect the lives of children and families. Nationwide, a movement to allocate public funds only to evidence-based programs is currently under way. Oregon recently passed legislation that restricts funds to proven-effective practices. And although the Washington Legislature did not pass a similar bill this past session, we expect

the issue to resurface next year.

And so we decided to use scientific methods to prove that our programs work. We didn't have the funding or staff time to undertake a large-scale experimental impact evaluation, but we could manage a small-scale project. We were certainly better positioned than most nonprofits to do so. We already had a team of directors and frontline staff that met monthly to work on evaluation. We were already collecting a lot of data about the children and families we serve. And we had our very own statistics expert, Susan Hautala, who took a yearlong sabbatical from her professorship at the University of Washington's School of Oceanography to help with the project.

With this team in place, we could conduct a more scientific evaluation. But even with more resources and support than most small nonprofits have, and with promising preliminary results, our evaluation couldn't tell us what we wanted to find out. We came to question the wisdom of policymakers who require that nonprofits scientifically prove their impact.

## Hazy Findings

Our research question was simple: Do children in the New Futures after-school program – which builds literacy skills, provides arts and enrichment activities, and connects parents and schools – improve their reading scores over the course of one academic year more than children not in the after-school program do? Because standardized reading scores were available, free, and simple to analyze, we chose to use this as our outcome measure.

To help plan our evaluation, we first talked to 25 regional and national experts. The King County Children



**LAURA SILVERSTEIN** (far left) is the associate director of New Futures. She has been with the Burien, Wash.-based organization since 1999. She directs New Futures' evaluation team and leads strategic initiatives.

**ERIN J. MAHER** is the director of program evaluation at Casey Family Programs, a foundation focused on serving children and families in the

child welfare system. She has served as an evaluation consultant for New Futures since 2005.

and Family Commission invested \$2,500 in the project. With these funds we hired Erin Maher, an established social policy researcher from the University of Washington, to consult on the project.

Armed with our consultant, we approached the Highline School District to see if they could give us test scores for both New Futures children and a matched comparison group. The district generously offered to pull one year's worth of scores and demographic information. We then had information on 465 kids who live in the apartment complexes we serve, 81 of whom were in New Futures programs. We knew their grade, gender, ethnicity, primary language, home language, and whether they were in special education or the English language learners program. We could use this information to compare New Futures children to similar, non-New Futures children.

We were thrilled when Susan brought the results to our lunch meeting. Looking at the raw data, we saw that New Futures children gained one-third of a grade more in reading than did children in the comparison group.

Our elation didn't last as long as our sandwiches, however. Susan next showed us the test for whether the difference between the scores was statistically significant – the mathematical gold standard for showing that a finding is not just a fluke. To our dismay, she said the differences between the two groups did not reach conventional levels of significance.

We had put our best resources – all our eggs – into this project, to prove for once and for all that organizations like ours can make a difference. And we failed.

What went wrong? Maybe we just didn't have enough participants for

our study. Relying on the school district's data constrained our sample size. Maybe we chose the wrong outcome measure: reading scores. After all, our program is geared toward holistic, long-term improvements, not targeted, short-term ones. Maybe the reading tests themselves are culturally biased, and therefore cannot detect the positive changes our program is making among our immigrant clients.

And, of course, it is possible that our programs don't work. But our internal evaluations suggest otherwise. Using a standardized test, our evaluation team found that New Futures children improved their oral reading skills 1.4 grades over a school year. Their parents are also benefiting, with 79 percent reporting increased involvement in their children's education, 83 percent feeling better able to meet basic needs, and 75 percent feeling more connected to their community. All of these factors have been linked directly to school success.

In the end, it would take a randomized control trial or some other sophisticated study to determine whether our programs work. But we are unwilling to deny services to people who need them the most – that is, to randomly assign some of our clients to a no-treatment condition. We also can't afford such a resource-intensive study.

### Endangering Innovation

To be "accountable," programs are supposed to be evidence-based. But organizations like ours do not have the resources to generate the evidence that funders and the public want. Indeed, small, flexible, community-based programs that serve diverse populations are among the worst-funded in the social sector.

Yet the kind of rigid, narrow

accountability that funders are demanding is of questionable validity. Scientific evaluations generally require staff to standardize interventions and deliver them consistently over long periods of time, regardless of individual needs, cultural considerations, or changes in circumstances. In contrast, New Futures aims to be flexible, innovative, and culturally competent. And so the very qualities that staff and families of New Futures believe make the program effective are the qualities that make measurement difficult.

Requiring nonprofit organizations to be accountable for their impact is, in many ways, a good idea. The children and families of our communities deserve programs that work. But they also deserve cutting-edge programs that respond continually to their changing needs. Organizations need flexibility and time to innovate before they are subjected to the rigors of the scientific method. Otherwise, programs will keep doing only what worked yesterday, instead of what works today. Funders and policymakers need to put more resources toward developing promising program models instead of shifting funds exclusively toward evidence-based programs. Otherwise, innovation will suffer for the cause of accountability.

At the Vintage Park apartments, 300 children continue to live in poverty. As winter approaches, they will come to our program without winter coats, in shoes that are too small. Their parents love them dearly, and have left all that is familiar to work impossibly hard to give them a better life. And New Futures will continue to support them with programs that they tell us work and that we think do, too, to the best of our ability to know. □