Title: **Golfing in a Hurricane: Education System Instability, Randomized Controlled Trials, and Children's Achievement**

Author: Robert F. Boruch, F. Joseph Merlino, and Andrew C. Porter

Date: June 28, 2012 3:30pm

Status: Working Paper CRESP 101JWRBEdit
Center for Research and Evaluation in Social Policy

University of Pennsylvania
Graduate School of Education
3700 Walnut Street
Philadelphia, Pennsylvania 19104-6216
215 898 0409 (office)
215 898 0532 (fax)
Contacts: robertb@gse.upenn.edu
andyp@gse.upenn.edu
jmerlino@21pstem.org

The research reported herein was supported by the US Department of Education, Institute for Education Sciences under IES award #R305C080009, The National Center for Cognition and Science Instruction. The opinions expressed are solely those of the authors and do not reflect the policies of the US Department of Education.

Table of Contents

Golfing in a Hurricane:

Education System Instability, Randomized Controlled Trials, and Children's Achievement

Robert F. Boruch, F. Joseph Merlino, and Andrew C. Porter

Date: June 25, 2012 3:30pm

Introduction

This report summarizes and extends ideas that have emerged from a large-scale cluster randomized trial on improvements to science curriculum for middle school children. We focus on instability in four school districts in which the trial was mounted, the implications for mounting further trials, and some implications beyond trials. The report expands on a commentary developed by Boruch, Merlino, and Porter (2012) for *Education Week*.

What are the Issues?

The immediate issue is that year-to-year instability and within-year instability of teachers who are involved in randomized controlled trials of education interventions is high in some, perhaps many, such trials. High-magnitude instability within and across years, one can argue, undercuts the possible effectiveness of multi-year interventions that depend on teacher stability, which in turn may affect children's achievement. In controlled trials on the effects of interventions that are tested well, this instability, we argue, can reduce the size of the intervention's apparent effect and the chance that the effects will be discerned. Put in other words by Merlino, “You can’t play golf in a hurricane.” Boruch, unlike Merlino and Porter, is innocent of golf, but nonetheless appreciates this characterization.

Consider further an issue that is important but subordinate to instability *per se*. It is that in published reports on the results of controlled experiments, instability related to the interventions has usually been labeled merely as “attrition from the study.” This scientific/statistical label is crude. It fails, for instance, to capture the different meanings and potentially far-reaching implications of instability in the design of interventions and in the design of experiments conducted to estimate their effects.

The instability phenomenon can be called “churn,” and one of us (Porter) favors this. This term, however, has different meanings in the national job market, and involves new hires and new separations that are regarded as healthy for an economy (Rampell, 2011). In what follows, we use the phrase ambient positional instability (API) to make the concern more precise in the larger context of experiments and in the contexts of turnover, churn, and mobility of personnel related to instability in schools or school systems.

What is the Magnitude of Local Ambient Positional Instability/Churn and its Importance?

The emphasis in what follows is mainly on API among teachers in schools within year and across year in randomized controlled trials in public schools. This is a choice that is constrained by resources for the moment. We recognize, for instance, that API at higher levels, such as principal and superintendent, is potentially as important, and we give references to evidence on this later. We also do not deeply consider mobility at lower levels, for instance, instability in cadres of tutors or students who switch schools, though this, too, is potentially important. See, for instance, Hanushek, Kain, and Rivkin (2004) on administrator stability and Mehana and Reynolds (2004) for meta-analyses of studies on student mobility. In what follows, we use illustrations from randomized trials in differing sectors to justify our interest in and our judgment about the import of ambient positional instability/churn/etc. in this context.

Since 2008, the National Center for Cognition and Science instruction has been conducting a large-scale cluster randomized trial on enhancing science curriculum in 182 middle schools. As mentioned at the outset, in this study approximately 42% of the Philadelphia middle school teachers across 92 of its schools who began participating in the intervention in the summer of 2009 were no longer able to participate in the intervention by September 2010. In 68 schools in two cities in Arizona, the average instability rate after one year of the intervention was about 24%, using as a base the number of teachers who consented to participate in the study. But the range in rates from school to school in each year is remarkable, at 1% to 62% (Michael Baker, personal communication, February 22 2012). There were no substantial differences in API between control and two intervention arms of the trial.

Other impact evaluations have generated similar evidence on instability per se and, by implication, the idea that the instability is a silent but potentially important factor in determining intervention effectiveness. For instance, in Garet et al.'s (2011) report on a cluster randomized trial of a professional development intervention for middle school math teachers, roughly 50% of teachers who were recruited into the intervention and control conditions in the randomized trial left the package of activity by the end of the second year.

In research on teacher incentives, Springer et al. (2010) reported results of a randomized trial in Nashville, TN, that was designed to learn whether high bonuses paid to teachers would result in higher achievement of children in mathematics. Of the 296 teachers engaged in this voluntary study, half left the study by the end of the third year. There was no difference in rates of leaving between control and intervention/bonus arms of the trial. Year by year rates of departure fluctuated but for reasons that were at least partly understood.

Hanson et al. (2012) reported on a cluster randomized trial involving 50 schools in a study designed to estimate the effect of a character education program on fifth-grade students' achievement. Between the year 0 in which random assignment was made and Year 1 of implementation, 18% of teachers who began in the study had moved out of the school, were assigned to an another ineligible grade, or withdrew from the study. The instability persisted from the first to the second year. About 23% of the teachers who had joined up for Year 2 of the study moved out of the study's ambit by the end of the year.

Consider further the Heller (2012) report on a randomized trial of a professional program for middle school science teachers in schools in California and Arizona. Of 181 teachers who were initially engaged and randomly assigned to the SCIENCE TM program to control conditions, "48 left the study before data collection was completed" (p. xi). This was over the period from Spring 2009 to Spring 2010. There was no appreciable difference in the rates of departure across the arms of the trial.

From Silverstein, Dubner, Miller, Glied, and Loike's (2009a) quasi-experimental study in New York, we learn that "only 54% of teachers who participated in the national SWEPT study's first year were available to participate in its second year" (p. 3). SWEPT involved science teachers in the city's public high schools.

Instability occurs not only in some U.S. experiments of course. In the first ever large-scale randomized trial in Italy, about 60% of teachers who began in the trial ceased engagement with it by the second year. The year-to-year instability for middle school teachers in Italy is about 35% but can reach over 80% for non-tenured teachers (Barbara Romano, personal communication, Sept. 13, 2011).

It may be of small comfort to education researchers that, in crime prevention research, API operates as well, though not necessarily at the service provider (cop) level. Police chiefs come and go, sometimes at an unnerving rate, in some U.S. cities, with the worst case being four chiefs over five years in one jurisdiction. Such instability at high levels is arguably important despite the inertia of complex organizational systems. See Boruch (in press) and the references therein.

We focus on ambient positional instability/churn in this report because its importance is not well recognized. However, we hasten to recognize that API/churn may *not* be an issue in some randomized trials. Consider, for instance, Weijekumar, Hitchcock, Turner, Lei, and Peck's (2009) report on a cluster randomized trial of the Odyssey Mathematics program for grade 4 students in a sample of schools in the mid-Atlantic region. It appears that *no* teachers disappeared from the trial during its course. Our first surmise was that the trial lasted only one year. In correspondence with one of the report's authors, we learned that he makes the same surmise. He emphasized that a second independent trial, of Connected Math, saw notable shifts in teacher position over the course of two years (Herbert Turner III, personal communication, June 1 2012). Similarly, we had surmised that stability in the Odyssey Math trial might be attributed to the fact that the sample of schools involved was suburban or exurban rather than urban. Again, one of the study's authors agreed with this surmise. He and we agree that deeper investigation of trials run under the auspices of the IES might shed further light on the matter.

Definitions and API beyond Local Jurisdictions

To put our definition of local API/churn into a broader context, consider Exhibit 1. Developed by Erling Boe (personal communications, April 16 and May 11 2012), it illustrates various definitions of attrition, turnover, and churn, and how they are related (or not) to one another in contemporary research. Boe's taxonomy is based mainly on categories developed to understand turnover in public schools in the United States at the national level rather than at local levels. The variations in definitions are complex, though Boe's taxonomy brings some order to the matter.

At present, our paper considers only within-year and across-year instability in the positions of teachers in local jurisdictions, in particular schools, teaching subject areas, and grade bands. It does not consider statewide or national evidence deeply. It does, at times, refer to instability at administrative levels inasmuch as this may influence instability at the teacher level. In particular, our paper focuses on categories I and IV in Boe's taxonomy, but it does so at local or sub-national levels, rather than the national level, and it considers within-year in addition to cross-year instability.

EXHIBIT 1

SOME DEFINITIONS RELEVANT TO PUBLIC SCHOOL TEACHERS (Boe, 2012)

I. Teacher Churn or Ambient Positional Instability (Boruch et al., 2012): Includes:

A. Attrition from teaching employment (called leavers by NCES)
B. School migration (called movers by NCES)
1. Within the same district
2. Between school districts, or to private schools
C. Teaching area transfer (Boe's term, operationalized as among 12 teaching areas)
D. Teaching level transfer (transfers among elementary, middle, and high-school levels)

II. Teacher Turnover (Boe et al.2008): Includes:

A. Attrition from teaching employment (called leavers by NCES)
B. School migration (called movers by NCES)
1. Within the same district
2. Between school districts, or to private schools
C. Teaching area transfer (Boe's term, operationalized as among 12 teaching areas)

III. Teacher Turnover (Ingersoll, et al.): Includes

A. Attrition from teaching employment (called leavers by NCES)
B. School migration (called movers by NCES)

IV. Teacher Retention (called stayers by NCES): Types include

A. Retention in teaching employment
B. Retention in same school district (includes retention in teaching employment)
C. Retention in same school (includes retention types IV A and B)
D. Retention in same teaching area (includes retention in teaching employment, but not necessarily in same school or same teaching level)
E. Retention at same teaching level (includes retention in teaching employment, but not necessarily in same school or same teaching area)

V. Teacher Retention: Used by Useem et al. (2007) for their Figure 4 [type IV. B., above]

VI. Levels of Analysis

A. National level (mostly SASS/TFS data by Boe and Ingersoll)
B. State level
C. District level (Philadelphia data by Boruch et al. and by Useem et al.)

In the literature on teacher turnover at the national level, “turnover” refers to the major changes in teacher’s assignments from one year to the next. “Turnover includes three components … leaving teaching employment (commonly called attrition), moving to a different school (commonly referred to as school transfer or migration) … and teacher area transfer” such as transfer from teaching one grade to teaching another (Boe et al., 2008, p. 8).

Definitions used at the local or state levels vary from the ones used at the national level. We judge this from personal communications with colleagues at the Pennsylvania Department of Education, among others. This non-uniformity in vernacular complicates efforts to understand the phenomenon. In this paper, we use the label ambient positional instability (API) to distinguish the phenomenon from conventional teacher turnover and to tie it to the context of randomized trials of particular interventions.

Reasons for Ambient Positional Instability

In Porter et al.’s (2012) trial, the reasons for API in Philadelphia include teachers “leaving the Philadelphia school (in which the experiment is embedded) on account of maternity leave, subject area reassignments (teaching math instead of science, for instance), grade band reassignments (teaching fourth grade rather than eighth grade), and other within school reassignments (going from being a teacher to becoming an assistant principal)” (Merlino, personal communication, March 25 2011). In the Arizona stratum of the trial, for instance, from one year to the next, teachers who for whom information is available left on account of transfers within school, transfers across school within district, transfers out of district, changes in teaching responsibility (subject or grade), and promotions into administrative positions (Baker, personal communication, February 22 2012).

The aforementioned Springer et al. (2010) report on a three-year trial to understand the effect of monetary incentives for teachers on middle school children's math scores says that teachers disappeared from the trial "most frequently, because they left the district, stopped teaching middle school mathematics—though they remained teaching in the middle schools—or moved to elementary or high schools in the district." (p. 9)

The Hanson (2012) report cited earlier, concerning a randomized controlled trial of a character education program, was conscientious in providing information on the teacher pipeline and reasons that teachers disappeared from the study. In the first year, for instance, 101 out of 117 teachers who left the study's ambit did so on account of being moved out of the school or assigned to an ineligible grade. The Year 2 results suggest that about 142 out of 615 left the study over the course of the year for similar reasons.

Silverstein et al.'s (2009a) quasi-experiment in the SWEPT program in New York reported that by the third year of their study, one teacher had left on account of maternity and two had been promoted to assistant principal. More important, 63 out of 95 eligible teachers left the study because they were reassigned to teach a non-Regents course, transferred to a different New York public school, lacked a non-participating teacher (control group member) in the same school teaching the same science course, or their school received a waiver from Regents exam administration (Silverstein et al., 2009b)

Sean Cavanagh (2011), writing in *Education Week,* covers another case in point. Based on a survey of the American Association of School Principals, he reports that 65% of school superintendents say they cut jobs in the last academic year, a plausible reason for API. For example: “A retired elementary school principal was replaced by a counselor. The counselor was replaced a middle school English teacher. The middle school teacher was replaced by a 5th-grade teacher. And on it went.” Elsewhere, a senior school official, a Mr. Kuhn in north Texas, referred to the “fruit basket” of reassignments attributed to budget cuts.

Garet et al.’s (2011) report on the second year of their middle school math professional development program trial contains no information on why teachers left the experiment, nor does it define categories of leaving. The report does, however, reiterate that teachers in each arm of the trial left at the same rates. From this fact, one may infer that teachers’ merely participating in one or the other arm of the trial had no effect on their departure rate.

Similarly, Heller (2012) provides evidence that the teachers who left their trial on middle school science professional development did not differ in character across arms of the trial. The report avers that “29 of the teachers (out of 48 who left the study) were not retained for reasons outside their control” (p. x).

The API Lacuna in Reports on the Results of Randomized Controlled Trials

At our request, Arellano (2011) reviewed all articles published in a first-rate peer-reviewed education research journal from March 2008 through September 2011 to learn whether and how the topic of instability was handled. Of a total of 46 articles, 29 reported on randomized controlled trials. Of the 34 reports on estimating outcomes of RCTs or QEDs, 19 reported no discernible and important differences in outcome between the control condition and the intervention condition(s). *None* of these 19 articles discussed instability in the work force in the intervention or control arms of the trials. In particular, none focused on instability as a potential explanation for the intervention's failure to produce discernible effects. Explanations apart from instability were put forward speculatively to explain *post facto* the absence of effect, e.g., weak implementation, poor fidelity, teachers' experiences, or irrelevance of the outcome measure.

There are several possible reasons that API is not well recognized in reports on the results of randomized trials. First among them is that the experiments reported in at least one mainline journal involve short time frames. In particular, 26 of 31 reports give a time frame involving only one year. API from one year to the next is irrelevant for such interventions, although within-year instability is relevant, as are temporary declines in outcome performance on account of new tasks that must be learned on the implementation/input side ("implementation dips"). The five published studies of work that extended beyond one year did not discuss instability over the period covered.

A second reason for not considering API in the context of controlled trials may be underlying assumptions or simple ignorance. For instance, many (perhaps most) scholars may not think instability is important in their controlled trials on interventions in school settings. Indeed, in informal conversations between Merlino and a principal investigator from a research firm that does good work, this investigator said, " I just consider it part what I have to deal with in the study, it is a given." The implication is that one need not regard the topic as a phenomenon worthy of study in its own right, and one might not study it if one is not paid to do so.

A third reason that API might not be considered in reports on the results of trials is that contemporary reporting standards, though they are good, do not require that one do so. For instance, the CONSORT Statement (CONSORT Group, 2011) on adequate reporting of clinical trials in health attends clearly to attrition from an experiment's pipeline. But there is no attention to its types or the reasons for it. The standards for reporting that the IES What Works Clearinghouse employs are based partly on CONSORT. The IES standards attend well to attrition/retention, but they don't attend to underlying reasons for instability. Experiments in the criminal justice sector attend to attrition, too. A special issue of the *Journal of Experimental Criminology* (Strang, in press) explores this more deeply attending to instability at the police chief and other levels.

Instability in Education Systems and Children's Achievement: Relation

We think it is reasonable to question the strength of the statistical relationship between what is understood about instability at all levels in school systems and the achievement of children in those systems. Contemporary evidence on this topic is choppy. There have been no controlled trials on this subject, as far as we know.

Consider, for instance, Fuller and Young's (2009) report on the tenure of middle school principals in Texas. The difference in leaving rate between new principals at low-performing schools and those at high-performing schools is about 8%. After three years, this instability rate is 46% in low-performing schools and 34% in high-performing schools. Whether principals leave on account of low-performing students or whether students perform poorly on account of mobility of principals cannot be determined from the data at hand. The effect on student achievement was not studied directly.

In a Michigan study, Keesler (2010) analyzed school level variation in achievement from eighth grade to 11th grade as a function of school level teacher retention rates. “The results of this study suggest that school level teacher retention rate is significantly related to school mean mathematics and ELA achievement, and that it explains 31% of variance in school mean mathematics achievement and 25% of the variance in school mean ELA achievement” (quoted in Stuckey, personal communication, May 2012). The statistical models employed were hierarchical. That is, they took into account school-, teacher-, and student-level characteristics including the usual resource-related control variables. The causal direction is unclear, of course.

Ronfeldt et al. (2011) depended on complex linear models to try to estimate the effect of turnover on ELA and math achievement of children in fourth and fifth grade in all New York City schools. Their data cover eight years and include adjustment/control variables such as children’s achievement in prior years, class, and school characteristics. They defined “turnover” in two ways. First, they considered the fraction of teachers who had left the school and grade level between time points t and t+1. The average over years and schools was about 13%, with a large range (from 0 to 50%). Second , they considered new teachers for each grade and school in each year. About one-tenth of new teachers in each school and grade were new each year. “Estimated coefficients for turnover were negative and significant for test scores in both ELA and math” (p. 14). The effects are statistically significant, we believe, mainly on account of the very large sample size. It appears that with a 25% turnover, one can expect a reduction of 2% of a standard deviation in test outcome, based on their models. It is not clear how well-specified the models are, though many ingredients are used and the results are provocative. The effect size estimates may appear small until one thinks about the effect size of aspirin use on heart attack rates, which is similarly “small,” to judge from controlled trials in that arena.

Of course, each time a school district decides to “reform,” or “restructure,” or to use the old but accurate word “reorganize,” there could be lots of API/churn at many levels in the system. For instance, Fullan (1991) speculated two decades ago that most interventions designed to improve curriculum and instruction would often involve an early dip in performance, inasmuch as the new program would require new skills and understandings. Further, he speculated that new teachers would be especially vulnerable in this inasmuch as they have other concerns that have to take priority, including learning classroom management. Journalists often report on reform initiatives, as in Philadelphia (Graham, 2012). Journalists are often not able to follow up on the reform’s consequences, including temporary instability that may induced by the initiative. That is up to us.

Educational Systems Instability More Generally, Available Evidence, and Beyond

Instability in an education system may occur at many levels: within nation, within a state or province or region, within a local school system or a school, and within a classroom. Of course, the changes in people—state or local system superintendents, principals, teachers, and students—may occur across year and within year. Changes can occur for a variety of reasons, as suggested by the cases described earlier.

Most important for the education sciences, empirical data on ambient positional instability/churn, may or may not be easily accessible at any of these levels. Nor may it be available for all these different kinds of people. Data on within-year instability is harder to access than are data on cross-year instability. Consider the following.

Background information from national studies seems important to put local jurisdictional data into context. For instance, recall that national data from the National Center for Educational Statistics’

Teacher Follow-up Surveys suggest that about 25.5% of teachers leave teaching employment during the first three years of (Boe et al., 2008). From Boe, Cook, and Sunderland (2009), we learn from the same data source that total annual teacher turnover, given by teaching field and year in 2004-2005, for all public education, was nearly 30%. This statistic includes attrition, teaching area transfer, and school migration based on unduplicated counts.

At the national level, empirical evidence on year-to-year turnover of teachers is available from probability sample surveys such as the Teacher Follow-up Survey (TFS) funded by the National Center for Education Statistics. Because the TFS is a national probability sample, estimates of turnover or other more nuanced indicators of change among teachers cannot be made for *local* schools or school systems. In particular, the national-level data can help in distal and weak benchmarking. See, for instance, Silverstein et al.'s (2009a) claims about how the SWEPT program enhanced teacher retention. But the national probability sample data are of no direct help in designing education interventions at the local level, nor are they of help in the statistical design of the experiments used to evaluate local interventions.

At the municipal level, case-based information is sometimes generated about instability, but its availability is choppy. Graham, Snyder, and Graham (2011), for instance, reported the following in a *Philadelphia Inquirer* piece on the city's school superintendent's departure from her position: "During her three years in Philadelphia, (Superintendent) Ackerman went through four chiefs of staff, three chief academic officers, and three communications officers." A report from ACTION United Education Fund (2012) in Philadelphia looks more deeply, finding that 35% of the highest-poverty schools got new principals in 2011-2012, as opposed to 15% of the lowest-poverty schools. ACTION also found that principals in the highest-poverty schools had an average of six years of experience, opposed to the nearly nine years of experience among principals in the lowest-poverty schools.

Studies of administrative data on this topic at the school district level are hard to come by at times. A recent exception is a Research for Action report by Useem, Offenberg, and Farley (2007) on Philadelphia schools. It tells us that of *new* teachers hired in 1999-2000, only 16% were in the same school in 2005). Of teachers hired in 2004-2005, only 68% were in the same school by the end of the second year of their tour. For the years 2002-2004, middle school teachers on average had the least experience (10 years), and high school teachers had the most (15 Years). See also Weinstein et al. (2009) on instability in newly created high schools in New York City.

At New York University, Tobias (2012) collected data on NYU graduates hired to work in New York City schools in 2006 and 2007. He reports that estimates of churn were about 27% after one year and 38% after two years in all schools and grades K-12. Marinelli's (2011) report for the Research Alliance for New York City schools gives data on all teachers in all schools from 2002 to 2009. About 45% of middle schools teachers left their schools within two years of entry, and 57% left within three years. Middle school teachers departed at a higher rate than others. See Marinelli (2011) for patterns and destinations of the teachers who left. The range of rates of departure within years across middle schools is remarkable, from a high of 66% in Manhattan to a low of 35% in Staten Island.

At the municipal school system level, it is possible that data on mobility of people within local schools or system turnover is being collected and archived for use in a database. For instance, we thought it possible that the Council of Great City Schools might maintain a database that characterizes the magnitude of turnover among teachers, principals, and system superintendents in the schools within the

Council's ambit. Such local data, if it can be accessed through the Council, could help understand ambient positional instability for experiments or quasi-experiments that are undertaken in those sites. Porter looked into the matter, and followed up in correspondence with a senior administrator at (Porter, 2011). The council does not have such information.

Regional studies could also help put local API into context. One of the few pertinent studies we have identified was done by Johnson, Huffman, Madden, and Shope (2011) at the Appalachia Regional Education Laboratory (REL). The authors' focus was on turnover of superintendents (not teachers) in Kentucky school districts. The REL report reinforces our own view that accessible instability data from year to year at the local or regional level are sparse in the particular case of school district superintendents as well as for other human resources. More to the point, over a quarter of Kentucky's school districts had three superintendents over the nine-year period of the study, or roughly one every three years. Nearly 50% had two superintendents over the nine years. According to the *Lexington Herald-Leader,* the Fayette County school district had two superintendents and two interim appointees over a five-year period before hiring the current superintendent (Xu, personal communication, October 2011).

Note also that we have not discussed small area estimators that might allow imputation from national to sub-national estimators. See Boruch and Terhanian (1996) for examples in education contexts.

Implications

The lines of thinking and evidence given in this report are choppy partly because evidence that can be brought to bear is itself choppy. Further, there has been little written on research policy or theory of instability that would help make thinking on this phenomenon more coherent and integrate it into education policy. Nonetheless, we are emboldened to offer implications that can be deduced from what we know now. We put the implications into question form as a matter of discretion rather than cowardice.

Site Selection for Randomized Trials. One implication of our concerns engenders a simple question: Should we try to do controlled trials in systems that are unstable? **F**or randomized trials, should only those sites that are stable rather than unstable be chosen if the intervention being tested presumes stability in the system? If one answers in the affirmative, this has further implications for generalizing results of the trial results. Obviously , tests of interventions that work in stable environments are useful, but those interventions may not work in unstable ones. The further implication for the design of controlled trials is that prior to a full-blown randomized trial, reconnaissance has to be done, and the sponsors have to understand that this pre-experiment reconnaissance has to be paid for. This line of thinking has important implications for research policy, of course.

Time Frame. If no schools are stable, should randomized trials be limited to testing those interventions with a nine-month duration at most (implementation dip included), rather than interventions that must be deployed over two or three years? The cross-year instability problem then disappears from the study. Of course, if one believes that education interventions must be deployed over longer periods to be effective, than some other options would have to be considered. And in fact, it is reasonable to suppose that some effective education interventions can be deployed only over two or more years.

Sequential Theory and Sequences of Trials. Should we build statistical designs and theories that yield estimates of effect that can be corrected for API or yield empirical indices that correct for API on theoretical grounds? Such model-based corrections are analogous to corrections for attenuation of correlations when measurements are subject to error.

This invites thinking about small theory in this context, which is better than no theory and better than complex theory. In particular, consider a shamelessly small and simple theory of how the effect size of an intervention in a trial may be degraded in moving from the laboratory or classroom trial, through efficacy trials, and then to eventual scale-up and deployment. Suppose that the intervention in the laboratory works to the extent that it moves children from the 50th percentile to the 65th percentile with a happy declaration that the effect size of one standard deviation differs from what one would expect from chance alone, i.e., statistical significance.

Assume further that the stability rate (1 minus the instability rate) of teachers in an efficacy trial is about .70. Further, suppose that in an effectiveness trial run under less than ideal and more realistic conditions, the stability rate is .50. Our simple theory says that effect size is degraded multiplicatively in moving from the lab to the field. That is, if we get 1.00 in the lab, we will at best get an effect size of (.70) (.50) (1.00) =.35 in the field. Let us be more realistic and suppose that principals and superintendents can move about also, and that their instability also affects the detectable effect sizes in a multiplicative fashion. If this happens, a stability rate of, say, .60 in the efficacy trial would result in an observed effect size of about .20. And so on.

To determine whether this simple and small theory holds up at all, one might exploit resources that accumulate data on various kinds of trials. These include the international Campbell Collaboration, the What Works Clearing House, the Coalition for Evidence-Based Policy, Slavin's Best Practice summaries, and others. None are currently set up to easily produce data that could easily be used to frame theory, but they have promise. The series of trials supported by the IES over the last decade of course constitute raw material for such an effort.

Consider further the typically weak linkage between laboratory efforts, in the cognitive science area for instance, and field trials. Laboratory efforts are often stable, as in the work by Jennifer Cromley at Temple University on ways to enhance children's understanding of science by using cognitive science principles to educate science teachers. Work by Hullleman and Cordray (2009) and their colleagues, building well beyond Boruch and Gomez's (1997) early work, is pertinent in trying to link the lab to the field. This engenders the question: what kinds of intellectual frameworks and statistical methods and indicators might be invented to do so?

Intervention Design and Development. Should we attempt to build interventions that dodge ambient instability or take it into account directly? For instance, peer-assisted learning and parent education can be construed as ways around the instability of teachers or principals, etc., that cannot be controlled by the school system. Such programs have met with some success. See, for example, reviews of controlled trials in this area produced by the international Campbell Collaboration at http://campbellcollaboration.org.

Or, one might assume that the instability is inevitable and uncontrollable, a fact of life. One might then build interventions that are prescriptive and routinely rectify gaps left by teachers or others who leave and accommodate new teachers or others entering the system. Consider, for instance, an analogy: Russian tanks versus German tanks during WWII. German tanks involved precision production, and the products required many tools for field repairs. In Russian tanks, crews were able to repair things that went wrong in the field with a couple of screwdrivers, wrenches, and a hammer. This is a grim but interesting metaphor that may be worth pushing in the domestic education sector. There may be better analogies for more pacific readers.

Third, and perhaps most important for randomized field trials, some people know enough, in Merlino's vernacular, to be ready and able to "plug the holes." For instance, when teachers disappear from any arm of a controlled trial, they must then be replaced by new teachers who themselves must be trained. Merlino recognized the vulnerability in this sensible contingency approach. New teachers involved in an intervention, for instance, may have many things to do—finding the restrooms, handling hallway mayhem, and coping with assistant principals Attila the Hun and Desdemona the Dreadful in the public schools and with Sister Mary Massacre in the Catholic ones. This is in addition to figuring out what the teacher is supposed to do to implement the new program being tested.

Control of API/Churn. Should we determine which dimensions of API are controllable in a given context and develop interventions that control it? For instance, one may think of introducing incentives or administrative systems that reduce API when that is warranted, such as "paying for staying," or induction and retention programs that others have suggested or tested. This, of course, requires some serious thinking about what level of API is desirable versus not desirable, quite apart from control devices. We have at least anecdotal evidence, from Lytle (personal communication, June 2012), for instance, that local administrative efforts can be made to understand productively what incentives matter for which teachers and how fungible resources can be used to actualize the incentives. Fullan and Hargreaves (2012) put the issue in more general terms of developing "professional capital," including incentives such as assuring economic returns in the teaching profession, and developing a culture/profession commitment, thorough preparation, well-networked and the capacity to make effective judgments. But research on API is sparse for local jurisdictions.

Concluding Remarks

We do not know which, if any, of the implications above are feasible or appropriate. They seem sensible to consider. Beyond the options we've laid out, we believe that it is important to begin to generate strategies and evidence on ambient positional instability/churn, on its plausible causes and consequences, on its control, and on ways to circumvent or dodge it. Absent attention to the matter, children will continue to be confronted by instability and arguably will be affected by it. More to the point, experiments on what works better for children that are well designed on statistical grounds will not be fruitful unless API is handled well.

References

ACTION United Education Fund. (2012). Revolving doors: *Findings from Philadelphia's highest poverty schools. Part 1: principal turnover and the importance of stable leadership.* Philadelphia, PA: Action United. Retrieved from http://www.actionunited.org

Arellano, E. (2011). Teacher turnaround as an issue for intervention success: A review of studies published in the *Journal of Research on Educational Effectiveness.* Philadelphia PA: University of Pennsylvania, Quantitative Methods Program, Briefing Document and Excel Spreadsheet (Email File).

Boe, E. E., Cook, L. H., & Sunderland, R. J. (2008). Teacher turnover: Examining exit attrition, teaching area, transfer, and school migration. *Exceptional Children, 75*(1), 7-31.

Boe, E. E., Cook, L. H., & Sunderland, R. J. (2009). *Trends in the turnover of teachers from 1991 to 2004: Attrition, teaching area transfer, and school migration* (Data Analysis Report No. 2007-DAR2). Philadelphia, PA: Center for Research and Evaluation in Social Policy, Graduate School of Education, University of Pennsylvania.

Boruch, R. F. and Gomez, H. (1977) Sensitivity, bias, and theory in impact evaluations. *Professional Psychology* (November), 411-434.

Boruch, R. F., & Terhanian, G. (1996). So What? The Implications of New Analytic Methods for the Design of NCES Surveys. In *The Futures Conference* (pp. 4-1 to 4-115). Washington DC: National Center for Education Statistics,

Boruch, R. (In press). Deploying randomized controlled trials in the interest of evidence-based crime policy*. Journal of Experimental Criminology.*

Boruch, R. F., Merlino, F. J., & Porter. A. (2012). Where teachers are replaceable widgets, education suffers. *Education Week, 31*(27), 20-21

Cavanagh, S. (2011). Budget pressures churn workforce. *Education Week*

CONSORT Group. (2011). CONSORT statement. Author. Retrieved from http://www.consort-statement.org/consort-statement.

Fullan, M. (2002). The implementation dip. *The Change Leader*, 59(8), 16-21.

Fullan, M., & Hargreaves, A. (2012, June 6). Reviving teaching with "professional capital." *Education Week*, *31*(33), 30, 36

Fuller, E., & Young, M. (2009). *Tenure and retention of newly hired principals in Texas*. Austin, TX: University Council for Educational Administration, Department of Educational Administration, University of Texas at Austin.

Garet, M. S., Wayne, A. J., Stancavage, F., Taylor, J., Eaton, M., Song, M., … Doolittle, F. (2011). *Middle school mathematics professional development impact study: Findings after the second year of implementation* (NCEE 2011-4024). Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Graham, K.A., Snyder, S., and Graham, T. (2011, Aug. 26). Ackerman and staff fell out of touch toward the end. *Philadelphia Inquirer*. Retrieved from http://philly.com

Graham, K. A. (2012, April 24). Phila. School District plan includes restructuring and school closings. *Philadelphia Inquirer*. Retrieved from http://philly.com

Hanushek, E., Kain, J., & Rivkin, S. (2004). Disruption versus Tiebout improvement: The costs and benefits of switching Sschools. *Journal of Public Economics, 88*(9-10)*,* 1721-1746. doi: 10.1016/S0047-2727(03)00063-X

Hanson, T., Dietch, B., & Zheng, H. (2012). *Lessons in Character impact evaluation: Final report* (NCEE 2012-4004). Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Heller, J. J. (2012). *Effects of Making Sense of SCIENCE™ professional development on the achievement of middle school students, including English language learners* (NCEE 2012-4002). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Hulleman, C. and Cordray, D. (2009) Moving from the lab to the field. *Journal of Research on Educational Effectiveness, 2*,88-110.

Johnson, J., Huffman, T., Madden, K., & Shope, S. (2011). *Superintendent turnover in Kentucky* (Issues & Answers Report, REL 2011–No. 113). Washington DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Appalachia. Retrieved from http://ies.ed.gov/ncee/edlabs

Marinelli, W. H. (2011). *The Middle School Teacher Turnover Project: A descriptive analysis of teacher turnover in New York City middle schools*. New York, NY: Research Alliance for New York City Schools.

Mehana, M., & Reynolds, A. J. (2004). School mobility and achievement: A meta-analysis. *Children and Youth Services Review, 26*(1), 93-119. doi: 10.1016/j.childyouth.2003.11.004

Porter, A., Merlino, J., Desimone, L., et al. (2012). Reports on the Cluster Randomized Trial on Enhancement of Science Curriculum for Middle School Math Students in Four Jurisdictions. AERA and others.

Rampell, C. (2011, Nov. 8). More churn in job market is hopeful sign. *The New York Times.* Retrieved from http://economix.blogs.nytimes.com/

Ronfeldt, M., Loeb, S., & Wycoff, J. (2011). *How Teacher Turnover Harms Student Achievement.* CALDER Report. . Washington DC: CALDER.

Silverstein, S. C., Dubner, J., Miller, J., Glied, S., & Loike, J. D. (2009a). Teachers' participation in research programs improves their students' achievement in science. *Science, 326*(5951), 440-442. doi: 10.1126/science.1177344

Silverstein, S. C., Dubner, J., Miller, J., Glied, S., & Loike, J. D. (2009b). Supporting online material for "Teachers' participation in research programs improves their students' achievement in science." *Science.* Retrieved from http://www.sciencemag.org/content/suppl/2009/10/15/326.5951.440.DC1/Silverstein.SOM.pdf

Springer, M. G., Ballou, D., Hamilton, L., Le, V., Lockwood, J. R., McCaffrey, D., … Stecher, B. (2012). *Teacher pay for performance: Experimental evidence from the Project on Initiatives in Teaching.* Nashville, TN: National Center on Performance Incentives, Vanderbilt University. Also at: http://www.performanceincentives.org

Strang, H. (Issue Ed.) (In press). Operational issues in randomized trials. *Journal of Experimental Criminology.*

Tobias, R. (2012). Re: Piece for the New York Times/Email to Andrew Porter, April 26 4:40pm and Email to Robert Boruch April 30 10:04am.

Useem, E., Offenberg, R., & Farley, E. (2007). *Closing the teacher quality gap in Philadelphia: New hope and old hurdles.* Philadelphia, PA: Research for Action. Retrieved from http://www.researchforaction.org

Weijekumar, K., Hitchcock, J., Turner, H., Lei, P., & Peck, K. (2009). *A multisite cluster randomized trial of the effects of Odyssey Math on the math achievement of selected grade 4 students in the Mid-Atlantic region.* (NCEE 2009-4068). Washington DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_20094068.pdf

Weinstein, M., Jacobowitz, R., Ely, T., Landon, K., & Schwartz, A. E. (2009). *New Schools New Leaders: A Study of Principal Turnover and Academic Achievement in New High Schools in New York City* (NYU Wagner Research Paper No. 2011-09). New York, NY: Institute for Education and Social Policy, New York University.

Acknowledgements

Torch Lytle, YueFeng Xu, Rui Yang, Daniel Stuckey, Rachel Skrlac, Ellen Donnelly, and Elias Arellano provided valuable research assistance or discussion or both in our work on this paper.