Listening. Learning. Leading.®

Research Report
ETS RR–12-17

What Can Repeated Cross-Sectional Studies Tell Us About Student Growth?

Russell G. Almond

Sandip Sinharay

October 2012

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

What Can Repeated Cross-Sectional Studies Tell Us About Student Growth?

Russell G. Almond and Sandip Sinharay

ETS, Princeton, New Jersey[1]

October 2012

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Associate Editor: Shelby Haberman

Reviewers: Yue Jia and Wendy Yen

Abstract

To answer questions about how students' proficiencies are changing over time, educational researchers are looking for data sources that span many years. Clearly, for answering questions about student growth, a longitudinal study—in which a single sample is followed over many years—is preferable to repeated cross-sectional samples—in which a separate sample is taken every year. Repeated cross-sectional studies, such as the National Assessment of Educational Progress (NAEP), however, are often readily available. Repeated cross-sectional studies conflate several sources of variability (differences in the initial status of individuals, individual differences in the growth curves, and individual-by-measurement-occasion differences) in ways that are not easily separated. Although repeated cross-sectional studies can provide information about the growth of the averages, the growth of the averages corresponds to the average of the growth curves only in very restricted circumstances. This paper reviews the literature on modeling growth with an eye to characterizing the limitations of repeated cross-sectional studies and understanding the sensitivity of the results to key decisions (particularly, choices of cut points). In most cases, repeated cross-sectional studies should be used to confirm and contextualize the results of more targeted longitudinal studies.

Key words: longitudinal studies, cross-sectional studies, growth models, measuring change, vertical scaling, causality

Acknowledgments

Henry Braun and Andrew Ho provided some valuable suggestions and references. Darlene Rosero provided assistance with the bibliography and references. Matthias von Davier, Andreas Oranje, Frank Rijmen, Mary Pitoniak, John Mazzeo, Helena Jia, Wendy Yen, Shelby Haberman, and David Williamson provided helpful feedback that improved the presentation of the paper.

The development of a longer version of this paper was supported by a special study contract from the National Center for Educational Statistics. The conclusions in this paper represent the scientific findings of the authors and do not reflect the opinions of the National Center for Educational Statistics, or ETS.

Above the first author's desk, there is a series of pictures featuring his twin daughters, taken every October in the same pumpkin patch (Almond, 2011). From the pictures, visitors can clearly see the (qualitative) growth that has occurred. However, producing a more quantitative measurement of the growth from this series of snapshots presents a number of problems. The pictures are shot with different zoom settings and cropping, so that the scale is different in each one. In addition, the girls are not always standing upright; sometimes one or the other is bending slightly in the picture. Despite these difficulties, the pumpkin pictures constitute a very informal *longitudinal* study. The twins are not identical, so each girl's growth can be tracked individually and aggregated to give an impression of average growth (provided that the various sources of measurement error can be accounted for).

Educators looking to study the growth of children in general will want to look at high-quality measurements of students that have been collected in a large number of years. Many such data sets, for example, the National Assessment for Educational Progress (NAEP), are *repeated cross-sectional studies* in which different individuals are sampled at regular intervals, rather than a longitudinal study in which the same individuals are measured at different points in time. A cross-sectional study is analogous to selecting a bunch of classroom photos at random from all of the fourth, eighth, and 12th grade classes in a single year. A repeated cross-sectional study selects a new group of photos of the fourth, eighth, and 12th graders every so often (every two years for NAEP). We have no idea if the same schools or individuals are represented in those photos. We do, however, have quite a lot of information about the schools and students written on the back of those photos (analogous to the NAEP background questionnaire data). For example, we may know the average income in the school district from which the photo comes. So what kinds of inferences is it possible to make about growth from this pile of photos?

Almost all researchers agree that longitudinal data is better than cross-sectional data for measuring growth (see, e.g., Singer & Willett, 2003, pp. 3, 9). A longitudinal study can control for individual differences in the students' initial statuses, along with other individual-level covariates and can study the variations in the patterns of growth, while repeated cross-sectional studies can only look at the growth of the average (which may or may not be the same as the average growth) or other summary statistics, and the growth measured is confounded with other uncontrolled components of variance between individuals and institutions. Bock (1991) pointed out, "Working with data from large-scale longitudinal studies has shown us, for example, that a curve based on

cross-sectional group means may be a very poor description of the actual course of growth seen in any given child." In fact, an estimate of a growth curve from repeated cross-sectional data is unbiased only when we assume that the growth function is linear (Willett, 1988). However, cross-sectional studies have the advantage of being less expensive (the children do not need to be tracked across years). In the case of NAEP, the data are already collected or scheduled for collection, so that the question is merely one of what can be learned about growth from additional analyses of the existing data.

Clearly, there are some things we can learn from the collection of school photos (our informal repeated cross-sectional study). For example, eighth graders are almost always taller than fourth graders. Other questions might be nearly impossible to answer from the collection of photographs; for example, are the taller eighth graders taller because they started out taller in fourth grade or because they grew faster (or a bit of both)? There are also some gray areas; for example, how do the growth patterns of girls and boys compare?

This paper presents a review of the literature on learning about change or growth from repeated cross-sectional data. This is an issue that arises not only in education, but in many other fields that make use of large government surveys, for example, economics, sociology, and epidemiology. Still, the most progress on this issue has occurred in the field of education. In particular, educational researchers have had a long-standing interest in measuring the effects of various instructional activities over time (Collins & Horn, 1991; Singer & Willet, 2003).

To make the literature more accessible to educational researchers, this paper is organized by research question. The following questions, which an educational researcher might want to ask of a repeated cross-sectional data source like NAEP, are addressed in the subsections below:

- What is the normal pattern of growth (growth rate) of students between two grades?
- Is a group of students on track?
- Do subgroups differ in their pattern of growth?
- How big is the effect of a policy on growth? Which policy is best?
- What is the value added by a particular educational program?

What Is the Normal Pattern of Growth of Students Between Two Grades?

With data from repeated cross-sectional studies, it is tempting to look at data from the same cohort at two different years (for example, fourth-grade scores from 2004 and eighth-grade scores from 2008) and ascribe the difference between them to growth. This simple comparison contains the hidden assumption that both assessments are measured on the same scale. Returning to the photograph metaphor, this requires that both photographs be *scaled* in the same way (i.e., the photographer is standing the same distance from the subject and the camera is set to the same zoom setting with the same size lens). The process of *vertical scaling* (Dorans, 2004; Kolen & Brennan, 2004) attempts to place the results obtained from each grade with different, but usually overlapping tests, on a common scale for comparison. There are various methods for vertical scaling and the choice of method can have an impact on the conclusions (Briggs & Weeks, 2009). In NAEP, vertical scaling is done only for reading and mathematics and only when the frameworks are established or revised for those subjects. After the initial scale is established (or reestablished), in each grade, each subject is linked back only to the previous administration (Gladkova & Xu, 2008). This process preserves the scale within an individual grade but does not automatically preserve the vertical scale necessary for cross-grade comparisons. In particular, each equating introduces a small error which may generate a cumulative error (or drift) over many equatings. This drift does not represent a big problem when making comparisons within grade between successive administrations, but it introduces a new and unmeasured source of error when making cross-grade comparisons. Note that constructing a proper vertical scale is also necessary to answer many of the more complex questions addressed later in this paper.

A related problem is the relationship between the year and the population variance. Many common statistical procedures assume *homoscedasticity*, that is, equal error variance throughout. However, depending on the procedure used for equating and the populations being studied, the variance may increase (or even decrease) with increasing grade (Ho, 2009). When the variance is changing, calculating effect sizes (observed effect divided by population standard deviation) requires some care so that the correct variance is used in the denominator.

Often when researchers ask about a *normal* pattern of growth, they are interested in growth trajectories, which will eventually lead to target levels of proficiency at the end of the student**s'** education. A better approach to answering those kinds of questions is to use the NAEP frameworks (or the equivalent for another repeated cross-sectional study) rather than the data. The NAEP

frameworks represent the ideas of a committee of experts about what the critical tasks are within a domain, and what the expected level of performance on those tasks is. Thus, the differences between the description of a proficient fourth grader and a proficient eighth grader provide a qualitative answer to the question, "What is the normal pattern of growth between two grades?" These definitions are fleshed out into more detailed achievement level descriptions as part of the standard-setting process. This leads naturally to the idea of tracking the proportion of students reaching each achievement level each year (Braun, 1988; Ho, 2007, 2008; Holland, 2002).

However, even this qualitative description of growth contains the implicit assumption that the expectations for fourth and eighth grade students remain constant across the years that the framework remains in force. In NAEP, the framework is revisited every so often (work on a new revision started in 2009). Although linking studies will try to provide a sense for how pre- and postrevision scores are related, it will be difficult to make cross-year comparisons.

Our recommendation is to use the testing frameworks to define a normal pattern of growth to proficiency. Even though this is mostly a theoretical framework, the empirical results of the repeated cross-sectional study play an important role in determining if real students follow the theory, and helping researchers understand when the theory needs to be refined.

Is a Group of Students On Track?

Answering this question requires first defining what is meant by *on track*. There are two ways to define on track: (a) the current status of the student is what is expected given the student's grade level and (b) the student is forecasted to meet some proficiency standard at a future time. The first approach really talks about change in status rather than growth; however, change in status is precisely what NAEP is designed to measure. The Basic, Proficient, and Advanced levels defined by the framework and the standard-setting process represent key mileposts in the expected developmental trajectory of the student. A natural choice is to define on track as being at or above one of those three levels, and the corresponding statistic to examine is the percentage of students falling at the chosen level or above.

While a single one of these *percentage above cut* (PAC) statistics is straightforward to interpret, care needs to be taken when looking at differences between them (Ho, 2008, 2009; Holland, 2002). In particular, small changes near the mode of the distribution of scores look larger than large differences near the tails of the distribution. Consider two populations of students: Population *a* is mostly high-achieving students, and the mode of their test scores is near the

Proficient level cut point. Population *b* is mostly low-achieving students, and the mode of their test scores is near the Basic level cut point. Now assume that an improved educational policy is applied to both groups of students, causing a small shift upward of the test scores of the students in both populations, by the same amount. At the Basic cut point, the PAC statistic will increase more for Population *b*, simply because more of that population is just below Basic, and the new policy pushes them over the line. Looking at the PAC (Basic) statistic (the percentage of students above the Basic cut point), it appears as if the gap between the two populations is closing. At the Proficient cut point, the opposite effect occurs. Because Population *a* has more students just below the cut point, its PAC (Proficient) statistic increases more than the corresponding statistic in Population *b*. To avoid getting fooled by statistical artifacts, Holland (2002) recommended looking at the PAC statistic for a selection of different levels.

Another way to define on track is to try and forecast whether the student will have reached a particular educational goal at a particular time (e.g., ready for the workplace or college upon graduation from high school). For the most part, this definition of on track requires longitudinal data to build the forecasts, so that repeated cross-sectional data cannot be used. Betebenner (2006) presented a rather interesting variant on this idea. Using state data from No Child Left Behind (NCLB) accountability tests, Betebenner calculated the chance that a student moves from a performance level in one time-point to another performance level in the next time-point (e.g., a move from Proficient in fourth grade in 2004 to Advanced in fifth grade in 2005). The values of these chances for the different performance levels can be combined in a mathematical framework to produce forecasts for the ability distributions in future years. Betebenner's analysis showed that either there needs to be a miraculous improvement in the transition probabilities (i.e., the schools must do a much better job of bringing stragglers up to speed) or NCLB's ambitious goal of 100% proficiency by 2014 will not be met. While estimating the full transition probability matrix requires longitudinal data, the repeated cross-sectional data does provide the margins of that matrix (that is, the percentage of students at each level in each year), which can be used to produce bounds on the growth rates (Tchen, 1980), although in practice those bounds may be very loose.

In summary, if on track means achieving a certain status, then repeated cross-sectional data can answer this question. However, we recommend looking at information about multiple points in the score distribution (e.g., at Basic, Proficiency, and Advanced levels) in order to provide a more complete picture of how the student populations are changing. If on track means forecasting future

growth, additional data sources are needed. In this case, the repeated cross-sectional data could provide bounds for the expected growth.

Do Subgroups Differ in Their Pattern of Change?

One of the more interesting uses for large-scale educational surveys like NAEP is monitoring gaps in achievement between different demographic groups. (This section considers natural groups such as those created by gender, race, income, or geographic location. The next section looks at the additional questions that arise when the groups are formed by different educational policies.) This is a particularly interesting problem because substantial increases in the proportions in the working-age population of the demographic groups with the lowest educational attainment will place strain on various social systems (Kirsch, Braun, & Yamamoto, 2007). Consequently, whenever there is a difference in achievement between a focal group (a lower-achieving subpopulation of interest) and a reference group (a historically higher-achieving group), an important question is, "Are the gaps between the focal and reference groups closing or widening?"

There are at least two different kinds of statistics that can be used to describe the differences between groups. The first is differences in the scaled scores (for example, differences in the average scores). The second is the differences in the PAC score. There are also two different kinds of comparisons we can make with repeated cross-sectional study data. The first compares the same grade levels at two different years (e.g., eighth graders in 2004 to eighth graders in 2008) in a *cohort-to-cohort* comparison. The second compares the same cohort measured at two different times (e.g., fourth graders in 2004 to eighth graders in 2008) in a *quasilongitudinal* study. Cohort-to-cohort comparisons aren't measurements of growth, rather they are descriptions of how the populations are changing over time relative to one another. Comparison of 2004 fourth-grade and 2008 eighth-grade samples look at the cumulative effect of all the changes in education and environment for preschool through eighth grade between 1990 (birth year for the 2004 cohort) and 2008. Some of what is observed in a quasilongitudinal comparison is change or growth; however, this is typically confounded with other factors. Nesselroade (1991) described three kinds of variability that can be present in temporal data:

- Individual-by-occasion—Small individual differences local to a particular measurement. This is the measurement error that the study designers try to minimize

but for which the variance must be estimated to calculate effect sizes and statistical significance.

- Individual-specific growth curve—Differences in the rates at which individuals change over time.
- Between-individual initial status—Differences in the starting values for the various measurements of the individuals.

The fundamental difficulty of a repeated cross-sectional study is that all kinds of variability are present in each measurement (especially after the first), and that the structure of the data collection provides no ways to apportion the observed variability among the three types. For that reason, most researchers studying change recommend only working with longitudinal data (e.g., Singer & Willett, 2003).

Differences in the Scaled Score Metric

When working in the scaled score metrics, the problem of comparability of the scales arises (Dorans, 2004; Kolen & Brennan, 2004). Returning to the photograph analogy, it is fairly simple to enlarge one photograph or shift its position up and down to match it to another and facilitate comparisons. However, if the papers the photos are printed on are slightly wrinkled from age, or if the photos were taken with a fish-eye lens that magnifies the center more than the ends,[2] the scales will contain subtle distortions. In this case, a difference of 10 scale points in the lower part of the scale may not be exactly equivalent to 10 scale points in an upper part of the scale. Such distortions arise naturally in educational assessments, which tend to gather the most information around the targeted difficulty of the items appearing on the test form. If the composition of the two tests differs (e.g., one has mostly easy items with a few hard items and one has mostly difficult items with a few easy items), then distortions are possible. Braun (1988) provided an example. Even if the equating procedure has done a good job of matching the difficulties, the standard errors can be different in different regions of the scaled score, complicating the interpretation of change scores (Yen, 1986).

In quasilongitudinal studies using the scaled scores, a natural statistic is the difference in scores between the first and second measurement; however, in order for that statistic to be interpretable, the two measurements must share a common scale (i.e., a vertical scale). A fundamental problem is that building a common scale requires common items. For example,

algebra is an important part of the curriculum for many eighth graders, but is presented in only a very anticipatory way to fourth graders. This marks a change in the dimensionality of the test, which can complicate the interpretation of difference scores (Martineau et al., 2008). A key problem is that parts of the construct that mark the difference between one group of students and another may be too easy or too difficult to provide much information about one of the groups, and hence the corresponding items may be dropped from the test specifications. For example, consider a mathematics test designed for first-through-third graders. Items requiring multiplication and division are likely to be too difficult for first graders (especially at the beginning of the school year) and therefore would be dropped from a first grade test as too hard. Items requiring simple addition should be easily solved by most third graders, and hence the items would be dropped from a third grade test as too easy. However, in order to measure mathematical progress through those grades, both addition and multiplication must be represented. That means that at least some of the students' testing time will be wasted with items that are either too easy or too difficult and hence provide little information.

Using the difference in the mean scale scores for each group as the criterion measure has one more important limitation. The mean only tells about what is happening at the center of the distribution. It may be of much more interest to look at what is happening to the students at the upper or lower ends of the scale. Holland (2002) recommended looking at the change in the scaled score for multiple percentage points in the distribution and summarizing that with a graph of change in score by percentile.

Differences in the PAC Scores

Another statistic that is readily available is the PAC score statistic. Unfortunately, due to a statistical artifact (illustrated in a previous section), the story told by the PAC statistic varies with the cut score, so that one cut point might show greater growth in the focal group and another greater growth in the reference group. (Ho, 2008, gave a dramatic illustration of this.)

This statistical artifact comes about because the choice of cut point is an arbitrary choice. A better description of the change would look at the difference at multiple cut points. Holland (2002) and Ho (2008) showed a number of graphical displays which provide a more complete picture of the differences in changes between two groups. One potential difficulty is that the PAC statistics are normally reported in the summary report only for the proficiency levels defined in the framework, and gathering a more complete picture may require the use of restricted-use data.

Braun (1988) introduced a trend statistic that avoids some of the pitfalls of looking at difference scores (see also Ho, 2007; Holland, 2002). This method yields the cut point at Time 2 that has the same number of students above the cut as the original cut point did at Time 1. This comparison should be done at various cut scores to give a complete picture of how the change effects the two populations at different points of the distribution. This procedure is similar to the procedure of equipercentile equating (Braun & Holland, 1982) and provides the most accurate results when the reliabilities of the scores at the two time-points are comparable. The implication is that scores at the two time-points do not need to be on the same scale. A graph showing the different cut points at Time 2 obtained using the approach of Braun (1988) from several original cut points at Time 1 is the same as a quantile-quantile (Q-Q) plot (Gnanadesikan, 1977).

Ho (2007) recommended a similar comparison, but in the inverse order so that probabilities instead of quantiles are compared. In particular, Ho computes the percentages of students that at Time 2 are below the cut point that had a fraction *p* of students below it at Time 1, for both the groups. This comparison is done on the probability scale instead of the scale of the assessment, and it is comparable to a probability-probability (P-P) plot (Gnanadesikan, 1977). Ho noted the similarity to receiver operating characteristic (ROC) curves, and that there was rich literature available for interpreting the scores. (The appendix to our paper provides mathematical details.) Holland (2002) described both methods in terms of the probability distribution function, with one method representing horizontal slices and the other representing vertical slices.

Because these methods compare quantities that are on the same scale, they avoid the difficulties, such as vertical scaling, that underlie difference scores, and are our recommended choice for between-group comparisons. Still, these methods are sensitive to the chosen cut scores or percentiles in the same way the PAC statistics are. Again, calculating these statistics at multiple points in the distribution is important for describing what is happening across the whole population of students.

How Big Is the Effect of a Policy on Growth? Which Policy Is Best?

This question is very similar to the last question, in that to answer it, the researcher must compare two or more different groups (groups of students who were educated under a given policy). All of the cautions and recommendations that apply to that problem apply to this one as well. However, another dimension is added to this question as presumably the researcher is trying to discover the causal effects of a policy. Repeated cross-sectional surveys, such as NAEP, are

almost always observational studies and the usual cautions about inferring causality from observational studies apply (Rosenbaum, 2002).

It is always difficult to determine whether an observed relationship is causal. Consider a study of the effects of school discipline policy on attendance. It is possible that a strict discipline policy might be put in place to prevent attendance problems (policy causes attendance) or in response to attendance problems (attendance causes policy). It is also possible that both attendance and discipline policy are based on community attitudes towards both discipline and attendance, which are almost impossible to measure in a school-based survey.

Changes in the composition of the underlying population are always a concern in repeated cross-sectional studies. Many factors (some related to education, some not) can affect emigration and immigration, which in turn will change the population size, the financial resources, and the demographic breakdown (particularly the distribution of educational attainment among the parents of school-aged children). As these factors could all contribute to changes in the distribution of proficiency across time, they hinder the estimation of the effect of a policy change.

The decision by a certain school, district, or state to implement a particular policy is never made in isolation. Certain pairs of policies may be nearly always implemented together, making it difficult to untangle the specific effects of a specific policy. Furthermore, if the policy effects are studied through aggregated units, choices in the weighting[3] may distort the effect sizes (Rosenbaum & Rubin, 1985; Wainer, Holland, Swinton, & Wang, 1985).

The decision to take up a policy is potentially based on many factors: economics, demographics, attitudes of the region about education, and politics. Often these variables are not directly measured in the study and could differ among the groups adopting differing policy choices. This is known as *selection bias*. When selection bias is suspected in an observation study, it is possible to perform sensitivity analyses to assess its impact on possible conclusions (Rosenbaum & Rubin, 1983). The basic idea of such an analysis is to model the selection in terms of an unobserved variable U that influences both the treatment assignment and the outcome variable; then one studies the change in the treatment effect corrected for selection bias as a function of the level of association between U and the treatment assignment—the change in the size of the effect due to a small change in the value of U will indicate the extent to which selection bias affects the conclusions.

Williams, Williams, Kastberg, and Jocelyn (2005) developed a method for studying causality when the policy variable and outcome variable may both affect each other. They build structural equation models with arrows pointing in both directions, but note that such models are identifiable only when an instrumental variable can be found that affects either the policy or the outcome, but not both. Even when such a variable can be found, caution needs to be used in the interpretation.

Some care must be taken when looking at causality with temporal data, as the critical factor may not be the current value of the variable but its value in the past (Gollab & Reichart, 1991). For example, the attitudes of a student's first-grade teacher about discipline may have as much effect on a fourth grader as the attitudes of the student's current teacher. Often, untangling these kinds of effects requires longitudinal data.

This is not to say that observational data play no role in the establishment or validation of causal conjectures (Shafer, 1996). First, when a relationship is apparent in an observational study, it is natural to investigate possible mechanisms for that relationship. Thus, an observational study may suggest causal conjectures and experimental or more detailed observational studies that would support or refute those conjectures. Second, if a causal mechanism is under consideration (say, from the results of a policy evaluation study), then it should be possible to predict how that mechanism would play out in the larger observational study. To the extent that the predicted observations are actually observed, an observational study provides additional evidence for (or against) the causal conjecture.

In educational policy research, true randomized experiments are very difficult to engineer. In this case, observational data is often better than no data at all. In the conclusions of their American Educational Research Association white paper, Schneider, Carnoy, Kilpatrick, Schmidt, and Shavelson (2007, p. 111) stated, "Through statistical techniques, large-scale [observational] datasets can approximate some of the probable causes and effects that experiments can establish more conclusively." They went on to observe, "Well-designed experiments maximize internal validity, whereas nationally representative observational datasets maximize external validity (Campbell & Stanley, 1963). Both are important." Rosenbaum (2002) provided a survey of relevant methods and appropriate cautions in interpreting the results of observational studies. If the goal is to make causal inferences, our recommendation is to synthesize evidence from a number of studies, some experimental and some observational.

What Is the Value Added by a Particular Educational Program?

Although much of the recent interest in value-added models stems from their use in accountability models (Sanders & Horn, 1994), the technique is, in fact, a much older (Bryk & Weisberg, 1976) way of measuring the effects of a particular educational program. The principal difficulty with measuring the effects of a program is that the students grow even under the control conditions. In a true experiment, the effectiveness of the program is measured by subtracting the growth observed in the randomized control group from the growth in the experimental (or treatment) group. A quasiexperiment (where the control and experimental groups are chosen for convenience) lacks the randomization to assure the researcher that the control and experimental groups are similar. Bryk and Weisberg (1976) suggested first building a forecasting model for predicting the posttest scores from the pretest and other background variables (in particular, the age of the subject). The *value added* by the program is then the difference between the forecast and what was actually observed.

Because value-added modeling involves building forecasting models, it is better suited to longitudinal data than repeated cross-sectional studies. In order to build the forecasting model, we wish to estimate the growth curve, a function that denotes the scores of a person at several time-points. What we would like is the average of the curves, where the average at each time-point is taken over the population of interest. However, the repeated cross-sectional data does not provide individual growth curves; the closest we can come to the average of the curves is the curve of average scores, which shows the average scores at several time-points. The curve of average scores will often not be an unbiased estimate of the average of the curves. It is unbiased only when *the growth curve* is a linear function. Willett (1988, p. 46) listed a number of interesting nonlinear models that get distorted by group averaging.

A second issue that needs to be carefully considered when building value-added models is that the initial status and growth are often correlated. Most researchers are aware of the regression effect. But in many cases, the correlation goes in the opposite direction, with the growth being positively correlated with the final status (Willett, 1988). This arises quite simply out of a linear growth model with unequal slopes. Assume all students' abilities in multiplication grow linearly but at different rates. Consider three students: Student *a,* who has a natural ability in mathematics, grows at a rate of 30 score points (on a particular test) per year; Student *b,* who has average mathematical ability, grows at a rate of 20 points per year; and Student *c,* who struggles with

mathematics, grows at a rate of 10 points per year. Now suppose that we are evaluating a third-grade mathematics curriculum and that the effect of that curriculum is that all students grow at their natural rate. The problem is that the zero point on the mathematics scale is not at the beginning of third grade—ideas about multiplication are introduced earlier. For the purposes of this example, we will say the beginning point of our multiplication scale is the beginning of first grade. Working this through we see that Student *a* will have a pretest score of 60 and a gain score of 30, Student *b* will have a pretest score of 40 and a gain score of 20, and Student *c* will have a pretest score of 20 and a gain score of 10. Apparently, there is a correlation between initial score and gain score; however, it is purely an artifact of the natural variation in the rate at which the three students learn mathematics. As the true zero point for the skills is usually in the distant past, this type of correlation is likely to appear whenever the effects of a particular educational policy are evaluated.

Finally, as value-added models are often used with observational studies, the usual cautions about inferring causality in such cases as stated in the previous section need to be kept in mind. In particular, one much discussed application is finding the value added by a particular teacher or school. However, the teacher effects are strongly confounded with the classroom effects defined by the cohort of students in that classroom and school effects are confounded with community effects for the community the school serves (Braun, 2005). Our recommendation would be not to try value-added modeling with repeated cross-sectional data; true longitudinal data is required to both build and verify the growth models that underlie the value-added analyses. Even so, care must be taken to ensure that all relevant covariates and possible confounding variables are properly accounted for in the model.

Concluding Remarks

In many respects, the use of repeated cross-sectional studies for measuring growth is like the use of observational data to study causality. Longitudinal studies are clearly superior to repeated cross-sectional studies for measuring growth because (a) there are certain sources of variability that are confounded in the cross-sectional study and (b) repeated cross-sectional studies provide an unbiased estimate of the average growth curve only when the growth curve is linear. As many states have instituted longitudinal data collection as part of their accountability systems, researchers increasingly have a choice between repeated cross-sectional data and longitudinal data.

It is important to keep in mind that repeated cross-sectional data measure change, not growth. Some of that change could be due to growth, but some also could be due to other factors. Table 1 summarizes some of the ways we recommend that repeated cross-sectional data should and should not be used to investigate student growth.

Table 1

Recommendations for Working With Data From Repeated Cross-Sectional Observational Studies

Don't	Do
Attribute all change between cohorts in a pseudo-longitudinal study to growth.	Remember that measurements in a repeated cross-sectional study have individual-by-occasion, individual-growth-curve, and between-individual variability.
Assume that tests given to different grades are on the same vertical scale.	Look at the definitions of cut scores for various proficiency levels, particularly, the claims they make about students.
Use curve of averages to estimate the average of curves.	Use Q-Q (Braun, 1988) or P-P (Ho, 2007) comparisons.
Assume that a difference between groups observed at one point in the distribution will be the same at all points of the distribution.	Compare groups at multiple cut points or percentiles.
Assume that two states/districts/schools that have chosen different policies are otherwise similar.	Perform sensitivity analyses to factors that could cause selection bias.
Assume that observed differences between groups are due to differences in policy.	Test findings made from studying observational data with experiments.

The advantage of targeted longitudinal studies is even stronger if the purpose of growth modeling is to estimate the effect of a policy. In large educational surveys, the policy indicator variables are often confounded with other policies and even the outcomes the policies are meant to address. For policy research, a small, well-designed, and targeted longitudinal study provides more evidence of the policy effect than large, general-purpose, repeated cross-sectional studies.

Nevertheless, a growth model can be used to make predictions about what should be observed in the repeated cross-sectional data, which should provide evidence to support (or refute) that model or policy choice. Furthermore, if the goal is to compare current and historical education

systems, there may only be cross-sectional data that covers the target time periods. The methods of Braun (1988), Holland (2002), and Ho (2007) in particular, can be used with cross-sectional data to look at how critical gaps in the educational system are changing, and these findings can then be used to suggest hypothesis that can be investigated with longitudinal studies.

References

Almond, R. (2011). *Pumpkins* [Photographs]. Retrieved from https://picasaweb.google.com/101504516516982228482/Pumpkins#

Betebenner, D. W. (2006, January). *Growth as a description of process*. Paper presented at the Festschrift dedicated to the life and work of Robert L. Linn, National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.

Bock, R. D. (1991). Prediction of growth. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 126–136). Washington, DC: American Psychological Association.

Braun, H. (1988). A new approach to avoiding problems of scale in interpreting trends in mental measurement data. *Journal of Educational Measurement, 25*(3), 171–191. Retrieved from http://www.jstor.org/stable/1434498

Braun, H. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Retrieved from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice, 28*(4), 3–14.

Bryk, A. S., & Weisberg, H. I. (1976). Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics, 1*(2), 127–155.

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand-McNally.

Collins, L. M., & Horn, J. L., (Eds.). (1991). *Best methods for the analysis of change: Recent advances, unanswered questions, future directions.* Washington, DC: American Psychological Association.

Dorans, N. J. (2004, May). *Vertical linking: Issues relevant to assessing growth*. Paper presented to the Evaluation Interest Group, Princeton, NJ.

Gladkova, L., & Xu, X. (2008). *Cross-grade scaling in NAEP: Linking design and analysis*. Unpublished manuscript.

Gnanadesikan, R. (1977). *Methods for statistical data analysis of multivariate observations*. New York, NY: Wiley.

Gollab, H. F., & Reichart, C. S. (1991). Interpreting and estimating indirect effects assuming time lags really matter. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 243–263). Washington, DC: American Psychological Association.

Ho, A. D. (2007). Discrepancies between score trends from NAEP and state tests: A scale-invariant perspective. *Educational Measurement: Issues and Practice*, *26*(4), 11–20.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*(6), 351–360.

Ho, A. D. (2009). A nonparametric framework for comparing trends in gaps across time. *Journal of Educational and Behavioral Statistics*, *34*, 201–228.

Holland, P. W. (2002). Two measures of change in the gaps between CDFs of test-score distributions. *Journal of Educational and Behavioral Statistics*, *27*, 3–18.

Kirsch, I., Braun, H., & Yamamoto, K. (2007). *America's perfect storm: Three forces changing our nation's future* (ETS Policy Information Report). Retrieved from http://www.ets.org/Media/Education_Topics/pdf/AmericasPerfectStorm.pdf

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Martineau, J. A., Subedi, D. R., Ward, K. H., Li, T., Lu, Y., Diao, Q., …Li, X. (2008). Non-linear unidimensional scale trajectories through multidimensional content spaces: A critical examination of the common psychometric claims of unidimensionality, linearity, and interval-level measurement. In R. W. Lissitz (Ed.), *Assessing and modeling cognitive development in school intellectual growth and standard setting*. Maple Grove, MN: JAM Press.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L. M. Collins & J. L. Horn (Ed.), *Best methods for the analysis of change* (pp. 92–105). Washington, DC: American Psychological Association.

Rosenbaum, P. R. (2002). *Observational studies* (2nd ed.). New York, NY: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcomes. *Journal of the Royal Statistical Society, Series B*, *45*, 212–218.

Rosenbaum, P. R., & Rubin, D. B. (1985). Discussion of “on ‘state education statistics’”: A difficulty with regression analyses of regional test score averages. *Journal of Educational Statistics*, *10*(4), 326–333.

Sanders, W., & Horn, S. (1994). The Tennessee value-added assessment system (TVAAS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation, 8*(3), 299–311.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W. H., & Shavelson, R. J. (2007). *Estimating causal effects using experimental and observational designs* (report from the Governing Board of the American Educational Research Association Grants Program). Retrieved from http://weraonline.org/uploadedFiles/Publications/Books/Estimating_Causal_Effects/Causal_Effects.pdf

Shafer, G. (1996). *The art of causal conjecture*. Cambridge, MA: MIT Press.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* London, UK: Oxford University Press.

Tchen, A. H.-T. (1980). Inequalities for distributions with given marginals. *Annals of Probability, 8*, 814–827.

Wainer, H., Holland, P. W., Swinton, S., & Wang, M. H. (1985). On “state education statistics.” *Journal of Educational Statistics*, *10*(4), 293–325.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422.

Williams, T., Williams, K., Kastberg, D., & Jocelyn, L. (2005). Achievement and affect in OECD nations. *Oxford Review of Education*, *31*(4), 517–545. Retrieved from http://www.jstor.org/stable/4618635

Yen, W. (1986). The choice of scale or educational measurement: An IRT perspective. *Journal of Educational Measurement, 23*, 299–325.

Notes

[1] Both authors conducted this research while employed by ETS. Russell Almond is now at Florida State University (ralmond@fsu.edu). Sandip Sinharay is now at CTB/McGraw-Hill (sandip_sinharay@ctb.com).

[2] In fact, many assessments are designed in this way, with a majority of items in the center of the difficulty distribution. This makes the test information function behave like a fish-eye lens, providing the most information in the center of the ability distribution.

[3] Producing an aggregate measure like an average involves taking a weighted combination of the individuals, although when simple random sampling is used, the weights are uniform. In more sophisticated analyses, and when complex sampling plans are used, the weights are chosen based on both the sampling scheme and the target population.

Appendix

Details of the Approaches of Braun (1988) and Ho (2007)

Braun (1988) introduced a trend statistic that avoids some of the pitfalls of looking at difference scores (see also Ho, 2007; Holland, 2002). This method starts with the cumulative distribution function (CDF), which is defined as $F(x) = P(X \le x)$, that is the probability that the score is below a given value *x*. The CDF should be nondecreasing in x. Note that $1 - F(x_0)$ is the percentage of students above a cut point x_0. It is also useful to think about the inverse CDF $F^{-1}(p)$, which is the value *x* such that $P(X \le x) = p$. (The inverse CDF can be difficult to calculate if the score is restricted to a finite number of discrete values. A common trick employed is to apply some kind of smoothing function to the CDF before inverting it. See Holland [2002] for details.)

Assume that we have measurements from two groups, *a* and *b*, and two different times, 1 and 2, and let $F_{1a}(\cdot)$, $F_{1b}(\cdot)$ $F_{2a}(\cdot)$ and $F_{2b}(\cdot)$ be the four different CDFs. The gap between Group a and Group b at Time 1 is $1 - F_{1a}(x_0) - (1 - F_{1b}(x_0)) = F_{1b}(x_0) - F_{1a}(x_0)$.

To compare the growth in Group a and Group b at a point x, Braun (1988) recommended comparing $F_{2a}^{-1}[F_{1a}(x)]$ to $F_{2b}^{-1}[F_{1b}(x)]$. This formula yields the new cut point at Time 2 that has the same number of students above the cut as the original cut point did at Time 1. This comparison should be done at various values of x to give a complete picture of how the change effects the two populations at different points of the distribution. Note the similarity of the form $F_{2a}^{-1}[F_{1a}(x)]$ to an equipercentile equating of $F_{1a}(\cdot)$ and $F_{2a}(\cdot)$ (Braun & Holland, 1982). The implication is that $F_{1a}(\cdot)$ and $F_{2a}(\cdot)$ do not need to be on the same scale. Also, the graph of $F_{2a}^{-1}[F_{1a}(x)]$ is the Q-Q plot of $F_{1a}(\cdot)$ against $F_{2a}(\cdot)$ (Gnanadesikan, 1977).

Ho (2007) recommended a similar comparison, but in the inverse order so that probabilities instead of quantiles are compared. In particular, Ho recommended comparing $F_{1a}[F_{2a}^{-1}(p)]$ and $F_{1b}[F_{2b}^{-1}(p)]$. These statistics compare the percentage of students that at Time 2 are below the cut point that had a fraction *p* of students below it at Time 1. This comparison is done on the probability scale instead of the scale of the assessment, and it compares to a P-P plot (Gnanadesikan, 1977).