Opportunities and Challenges in Collecting and Analyzing Longitudinal Educational Data

David Kaplan
Department of Educational Psychology

Keynote Address
Israel Academy of Sciences and Humanities
Jerusalem, Israel
Portions of this talk appear in


Opportunities of using longitudinal educational data for policy analysis

- Growth curve modeling for continuous variables over discrete time.
- Growth mixture modeling.
- Stage-sequential change.

Challenges in collecting and using longitudinal educational data for policy analysis

- File matching
- Privacy

Conclusions: Extended Opportunities
Let’s start with opportunities.

Access to large scale, nationally representative longitudinal data now allow applications of sophisticated studies of growth and change over time.

Examples: ECLS-K, LSAY, ELS

ECLS-K (NCES, 1998)

- Arguably the most comprehensive survey of the development of academic competencies over the early elementary and secondary school years.
Early Childhood Longitudinal Study


(From NCES website)

- The children came from both public and private schools and attended both full-day and part-day kindergarten programs. They came from diverse socioeconomic and racial/ethnic backgrounds. Also participating in the study were the children’s parents, teachers, and schools.

- Children, their families, their teachers, and their schools provided information on children’s cognitive, social, emotional, and physical development. Information on children’s home environment, home educational activities, school environment, classroom environment, classroom curriculum, and teacher qualifications also was collected.
The ECLS-K followed the same children from kindergarten through the 8th grade.

The ECLS-K was designed to provide comprehensive and reliable data that can be used to describe and to understand better children’s development and experiences in the elementary and middle school grades, as well as how children’s early experiences relate to their later development, learning, and experiences in school.

The data collected across the years allow researchers and policymakers to study how various child, home, classroom, school, and community factors at various points in children’s lives relate to cognitive, social, emotional, and physical development.
Change in Continuous Variables over Discrete Time
Level – 1 (intra-individual differences)

- These differences are what we see in the graph: Children start at different levels, change at different rates, and end at different levels.

- Statistical models for intra-individual differences provide an estimate of the starting point and the rate of growth.

- Extensions of these models allow for growth to be non-linear.
Level–2 (Predictors of individual variability in growth)

- We are interested in the predictors of the variability we see in the individual growth curves.

- We can examine the impact of policy relevant variables on the average starting point and rate of growth over time.

- These predictors can even interventions designed to impact the growth parameters.

Level–3 would allow individuals to be nested in schools.
Growth Curve Model Diagrams

Figure 8.3  Initial growth curve model of science achievement.
Figure 8.4 Growth curve model of science achievement with time invariant predictor.
Figure 8.5 Growth curve model of science achievement with time invariant and time varying predictors.
Figure 8.6 ALT(1) model of science achievement.
Figure 1. Path diagram of sequential growth curve model for finger use and number combinations. "Fingers" stands for frequency of finger use and "Accuracy" stands for performance on number combinations. Note that for ease of reading, correlations among growth parameters are not presented but are estimated in the analysis.
Example of GCM: Kaplan (2002)

- Data utilized for this example consist of the kindergarten base year (Fall 1998/Spring 1999) and first grade follow-up (Fall 1999/Spring 2000) panels of ECLS-K.

- Only first-time public school kindergarten students were chosen for this example.

- The sampling design of ECLS-K included a 27% sub-sample of the total sample at Fall of first grade in order to reduce cost.

- The sample size for this example is 3575.

**Caveat:** This is an example, and there is no presumption that the results are substantively important.
Measures

- ECLS-K reading assessment measuring the range of basic skills up to reading comprehension.
- Indicator of full-day versus part-day kindergarten attendance.
- Measure of SES.
- Child’s age of entry into kindergarten.
Research Questions and Results

1. What is the general level and rate of growth in reading proficiency among the sample of ECLS-K children?

2. How do growth rates differ for children attending full-day and part-day kindergarten programs?

3. Does the effect of full-day versus part-day kindergarten programs change when controlling for the age of entry into kindergarten and SES?
Example of GCM

Stage-Sequential Change

Challenges

Conclusions: Extended Opportunities
Introduction

Opportunities

Growth Curve Modeling

Example of GCM

Growth Mixture Modeling

Stage-Sequential Change

Challenges

Conclusions: Extended Opportunities

Reading IRT Score vs. ECLS-K Waves

- Full w/ ses and age
- Part w/ ses and age

1 2 3 4

0 10 20 30 40 50 60

ECLS-K Waves
Let's consider the growth curves for the ECLS-K one more time.

Notice that there are clusters of different curve shapes.
Our concern with conventional growth curve modeling is that it presumes a “one-size-fits-all” model for the growth trajectories.

In other words, we are assuming that the growth trajectories can be described by a single average curve.

What if this is not true? What if there are distinct curve shapes?

The implications for interventions are important. Knowing if there are sub-populations with distinct curve shapes permits tailored interventions.

We need a method to discern if there are distinct growth trajectories, how many of them exist in the data, and how many individuals can be classified into having these shapes.

This is the goal of growth mixture modeling.
Growth mixture modeling is a type of general mixture model.

The idea is that in any given population, there might be mixtures of sub-populations of varying number and size.

The is a general idea that can be used in other statistical models, such as regression. The sub-populations are characterized by different model parameters.

For growth mixture modeling, the sub-populations are characterized by qualitatively (or substantively important) growth curve shapes. The model parameters here represent the shapes of these trajectories within sub-populations.
Introduction

Opportunities

Growth Curve Modeling

Example of GCM

Growth Mixture Modeling

Stage-Sequential Change

Challenges

Conclusions: Extended Opportunities

---

**Reading IRT Score**

**ECLS-K Waves**

- **Class 1:** Normal developing (72%)
- **Class 2:** Slow developing (15%)
- **Class 3:** Fast developing (13%)
Introduction

Opportunities

Growth Curve Modeling

Example of GCM

Growth Mixture Modeling

Stage-Sequential Change

Challenges

Conclusions: Extended Opportunities
Introduction

Opportunities

Growth Curve Modeling
Example of GCM

Growth Mixture Modeling

Stage-Sequential Change

Challenges

Conclusions: Extended Opportunities

---

![Graph showing reading IRT score across ECLS-K Waves for different groups with and without SES and age adjustments.](image)
Stage-sequential change: Kaplan (2005, 2008); Kaplan et al. (2009)

- Stage theories of development (e.g. Piaget, Kohlberg)
- ECLS-K also provides measures that can be used to test hypotheses regarding stage sequential change in reading competencies.
- Prior stage sequential studies of reading
  - Chall (1995)
  - Juel (1988)
    - Neither used advanced models for estimation
ECLS-K provides transformations of continuous IRT scores into probabilities of proficiency as well as dichotomous proficiency scores, which are used in this study.

ECLS-K instrument design formed clusters of reading assessment items having similar content and difficulty.

A child was assumed to have passed a particular skill level if he/she answered at least three out of four items in the skill cluster correctly. A fail score was given if the child incorrectly answered or did not know at least two items within the skill cluster.
We utilize the subtests of the ECLS-K reading assessment.

The measures used in this example consist of a series of reading assessments designed to measure basic skills that Whitehurst and Lonigan (2002) have identified as particularly salient in the first two years of school.

Specifically, the reading assessment Markov chain models for developmental processes 6 yields scores for (1) letter recognition, (2) beginning sounds, (3) ending sounds, (4) sight words, and (5) words in context.
In addition to the reading scale scores, ECLS-K provides transformations of these scores into probabilities of proficiency as well as dichotomous proficiency scores, which are used in this example.

To calculate dichotomous proficiency scores, the ECLS-K instrument design formed clusters of reading assessment items having similar content and difficulty.

A child was assumed to have passed a particular skill level if he/she answered at least three out of four items in the skill cluster correctly. A fail score was given if the child incorrectly answered or did not know at least two items within the skill cluster. In the case of exactly two items correct, a pass/fail score was given if the pattern of passes and fails for remaining proficiencies yielded could suggest an unambiguous pass or fail.
Table 3. Response probabilities and class proportions for separate latent class models

<table>
<thead>
<tr>
<th>Latent Class</th>
<th>LR&lt;sup&gt;b&lt;/sup&gt;</th>
<th>BS</th>
<th>ES</th>
<th>SW</th>
<th>WIC</th>
<th>Class Proportions</th>
<th>$\chi^2_{LR}$ (29df)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAK&lt;sup&gt;c&lt;/sup&gt;</td>
<td>.47</td>
<td>.02</td>
<td>.01</td>
<td>.00</td>
<td>.00</td>
<td>.67</td>
<td>3.41</td>
</tr>
<tr>
<td>EWR</td>
<td>.97</td>
<td>.87</td>
<td>.47</td>
<td>.02</td>
<td>.00</td>
<td>.30</td>
<td></td>
</tr>
<tr>
<td>ERC</td>
<td>1.00</td>
<td>.99</td>
<td>.98</td>
<td>.97</td>
<td>.45</td>
<td>.03</td>
<td></td>
</tr>
<tr>
<td>Spring K</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAK</td>
<td>.56</td>
<td>.06</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.24</td>
<td>4831.89*</td>
</tr>
<tr>
<td>EWR</td>
<td>.99</td>
<td>.92</td>
<td>.63</td>
<td>.05</td>
<td>.00</td>
<td>.62</td>
<td></td>
</tr>
<tr>
<td>ERC</td>
<td>.00</td>
<td>.99</td>
<td>.99</td>
<td>.96</td>
<td>.38</td>
<td>.14</td>
<td></td>
</tr>
<tr>
<td>Fall First</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAK</td>
<td>.52</td>
<td>.08</td>
<td>.01</td>
<td>.00</td>
<td>.00</td>
<td>.15</td>
<td>11.94</td>
</tr>
<tr>
<td>EWR</td>
<td>1.00</td>
<td>.92</td>
<td>.71</td>
<td>.05</td>
<td>.03</td>
<td>.59</td>
<td></td>
</tr>
<tr>
<td>ERC</td>
<td>1.00</td>
<td>.99</td>
<td>.98</td>
<td>.98</td>
<td>.42</td>
<td>.26</td>
<td></td>
</tr>
<tr>
<td>Spring First</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAK</td>
<td>.19</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.04</td>
<td>78.60*</td>
</tr>
<tr>
<td>EWR</td>
<td>.98</td>
<td>.90</td>
<td>.79</td>
<td>.35</td>
<td>.00</td>
<td>.18</td>
<td></td>
</tr>
<tr>
<td>ERC</td>
<td>1.00</td>
<td>.99</td>
<td>.98</td>
<td>.99</td>
<td>.60</td>
<td>.78</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> Response probabilities are for passed items. Response probabilities for failed items can be computed from $1 - prob$ (mastery).

<sup>b</sup> LR = letter recognition, BS = beginning sounds, ES = ending sounds, SW = sight words, WIC = words in context.

<sup>c</sup> LAK = low alphabet knowledge, EWR = Early Word Reading, ERC = early reading comprehension
Table 4. Transition probabilities from Fall kindergarten to Spring First grade.

<table>
<thead>
<tr>
<th>Wave</th>
<th>LAK</th>
<th>EWR</th>
<th>ERC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall K</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAK</td>
<td>0.30</td>
<td>0.69</td>
<td>0.01</td>
</tr>
<tr>
<td>EWR</td>
<td>0.00</td>
<td>0.66</td>
<td>0.34</td>
</tr>
<tr>
<td>ERC</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wave</th>
<th>LAK</th>
<th>EWR</th>
<th>ERC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall First</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAK</td>
<td>0.59</td>
<td>0.40</td>
<td>0.01</td>
</tr>
<tr>
<td>EWR</td>
<td>0.00</td>
<td>0.82</td>
<td>0.18</td>
</tr>
<tr>
<td>ERC</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Wave</th>
<th>LAK</th>
<th>EWR</th>
<th>ERC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring First</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LAK</td>
<td>0.30</td>
<td>0.48</td>
<td>0.22</td>
</tr>
<tr>
<td>EWR</td>
<td>0.01</td>
<td>0.13</td>
<td>0.86</td>
</tr>
<tr>
<td>ERC</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Goodness-of-fit tests

\[ \chi^2 (1048528 \, df) = 12384.21, \, p = 1.0 \]

\[ \chi^2_{LR} (1048528 \, df) = 6732.31, \, p = 1.0 \]

BIC = 44590.80

*LAK=low alphabet knowledge, EWR=Early Word Reading, ERC=early reading comprehension*
What are the challenges in building longitudinal data files?
What are the challenges in building longitudinal data files?

- We consider two operational challenges when collection and building longitudinal data.
What are the challenges in building longitudinal data files?

- We consider two operational challenges when collection and building longitudinal data

  - File matching
What are the challenges in building longitudinal data files?

- We consider two operational challenges when collection and building longitudinal data

1. File matching
2. Privacy
Can we combine data from various sources to provide a richer set of variables for research and policy analysis?

This problem is more or less straightforward depending on the level of aggregation.

Example: Additional country level data from the OECD can be added to PISA country level data for purposes of studying cross-country growth.

What about within Israel? What about schools? Students?
Creating synthetic longitudinal data

Example: ECLS-B added to ECLS-K

Requires statistical matching methods

- Matching individuals from different databases requires sufficient background information for the match
- Statistical methods are needed to check the validity of the match
- Knowledge of super-exogenous changes in the policy context is required
- Thought given to the population that the synthetic cohort now represents
Some Imputation Methodologies

- A small list of statistical matching and imputation methodologies:
  1. Nonparametric Hot Deck Matching.
  2. Stochastic regression imputation.
  3. Predictive mean matching.
  4. Bayesian linear regression via chained equations.
  5. Bayesian bootstrap predictive mean matching.
  6. EM bootstrap- hybrid Bayesian and EM with and without priors.

- These are all available in the open source software “mice” within the R computing environment.
Validating Imputation Methods

For any rigorous data fusion, validity criteria must also be established.

We consider the work of Rässler (2002) and her colleagues.

1. First Level Validity: Preserving individual values
2. Second Level Validity: Preserving joint distributions
3. Third Level Validity: Preserving correlation/covariance structure
Case Study From Iceland: Kaplan & McCarty (2013)

- 142 schools participated in either the TALIS survey or the PISA survey.

- Of these, 122 PISA and TALIS schools were able to be matched.

- The 20 schools that were unmatched were eligible for TALIS or PISA, but not both.

- An additional 39 schools were excluded due to large amounts of missing data on variables needed for the fusion procedures. Finally, 5 schools were excluded because they were identified to be influential outliers.

- Thus, the statistical matching procedures utilize data from 78 schools in Iceland with full information from the PISA and TALIS data sets.
Matching Variables

- School sector;
- The size of the school community; the total enrollment in the school;
- A measure of the availability of school material resources;
- The extent to which teacher absenteeism interferes with student learning;
- A measure of the extent to which student-related factors affect the school climate;
- A measure of the disciplinary climate of the school.
Unique Variables

From PISA:

Enjoyment of reading; summarizing skills. Both measures are averaged to the school level for analysis.

From TALIS:

Teacher job satisfaction; teacher self-efficacy. These measures were also averaged to the school level for analysis.
Figure 1: Kernel Density and QQ Plots For Bayesian Bootstrap Predictive Mean Matching
Challenges with Data Fusion

- What are some of the challenges with data fusion?
  1. Multilevel data fusion?
  2. Sampling weights – what is the relevant population?
  3. Temporal concerns. When were surveys obtained?
Opportunities with Data Fusion

Numerous opportunities await more data fusion.

1. Cost savings?
2. Greater coordination and collaboration in national data collection efforts?
3. Potential for great policy relevance - particularly when data files are matched with other national system indicators.
Privacy

- Obviously not a simple issue.

- Necessary to convince stakeholders of the security of the information – particularly information collected on children.

- How is this handled by NCES?

  - Two versions of the databases are available: Public-use and Restricted-use

  - Public-use data are freely available and have identifying information removed, making it possible to identify individuals, but also small areas

  - Restricted-use data is available under a license that specifies how the data are to be used and stored.

  - Strict penalties are imposed if random spot-checks reveal non-compliance with the license.
Just a few of the requirements

1. The Security Plan form must specify the exact location where the data will be secured. Only persons listed on the License can have access to the locked project office.

2. If you plan on using the data at a second location that is not on the same floor in the same building as your original secure project office, you will need to apply for a second License.

3. Restricted-use data, in any form, may not leave the secure project office or be moved off of the standalone (non-networked) desktop computer. All original CDs of the data must be kept under lock and key within the secure project office. The data must not be placed on a network server, laptop computer, external hard drive or USB thumb drive (memory stick).

4. Storing or using restricted-use data at your home is not allowed. The data must only reside within the secure project office at your institution as specified in your Security Plan form.
The challenges of mounting a longitudinal study and providing the data are real and serious.

However, the depth of what can be learned makes the effort worthwhile.

What other extensions are possible?
Large-scale cross-sectional and/or longitudinal surveys cannot measure every interesting variable.

While there may be theory driving the measurement of critical outcomes (e.g. reading literacy), not all cognitive and non-cognitive outcomes nor can all theoretically driven predictors be measured.

Response burden can be mitigated through various matrix sampling (rotation) designs.

There may be no new data collection in Israel. Extant data will be merged.

Perhaps it is better to thinking of these surveys as measures of important indicators of the health of the national educational system.
What are the characteristics of indicator systems?

1. They should be theoretically or empirically linked to other indicators.

2. They should serve to monitor real or simulated changes in outcomes of interest.

3. They should act as proxies to real policy instruments.

4. Their validation should come from movements in relation to the movements of other indicators, if such relationships are presumed.
Variables are linked to each other through the theoretical frameworks.

These variables are not usually viewed as indicators but rather as research variables.

What if they were conceived of as monitoring variables?

Proxies to real policy instruments suggests malleability.

Movement in relation to movements in other indicators suggests prediction.

Prediction, in turn, could lead to forecasting.
If interest centers on forecasting and prediction, then adherence to an idiosyncratic theoretical model is less important (and counterproductive) than focusing on combining many possible models to attain optimal prediction.

Quoting Hoeting et al. (1999)

“Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are.”(pg. 382)

Frequentist and Bayesian approaches to combining models exist and have been studied in the statistical literature.
A key characteristic of statistics is to develop accurate predictive models (Dawid, 1984).

A given model is to be preferred over other competing models if it provides better predictions of what actually occurred.

Longitudinal data provide a unique opportunity to build and test predictive models that can be tested by forecasting beyond a specific time point, and then see how the prediction matches the outcome of interest when that time point arrives.

Prediction/forecasting models can then be “calibrated” based on the accuracy of the predicted outcome.

For Israel, matching data files should be done with an eye toward which indicators are needed for policy purposes and predictive modeling.
Our research group is exploring frequentist and Bayesian approaches to model averaging for the purposes of building optimal predictive models in education.

We are examining these tools for both cross-sectional international data (specifically PISA and TIMSS), as well as national trend data (NAEP) and national longitudinal data (ECLS).

Although we understand the challenges of cost, burden, and privacy, our hope is that the opportunities for a “deep dive” into growth and development that also serves to monitor overall trends can be realized.
Thank You