A Simulation Model of the Black-White Reading Gap

Linwood Tauheed, PhD.

Associate Professor of Economics

University of Missouri Kansas City

December 11, 2025

[1]

[1] The author has received no financial support for this research and has no financial relationships or other potential conflicts of interest that apply.

ABSTRACT

We develop a System Dynamics model of the kindergarten classroom process leading to the disparity in Black-White letter recognition proficiency. We model teacher instruction in letter recognition in interaction with a learning-forgetting model of student learning.

Teacher expectations are operationalized in a “Teaching Curve” parameterized by teacher’s "intensity of instruction" and "proficiency goals", the levels of which are modified by the race and class of students.

The student’s learning process is modeled as a Learning Curve parameterized by non-raced/classed rates of learning and forgetting.

Model parameters (teacher’s intensity of instruction and proficiency goals and student’s rates of learning and forgetting) are calibrated using data from the Early Childhood Longitudinal Study - Kindergarten sample (1998).

Simulation results find that student's race and class, when taken as variables affecting teacher expectations, but not as variables affecting student’s cognitive learning process, can explain Black-White differences in kindergarten letter recognition proficiency.

A number of innovations in educational research are employed:

1. Modeling ECLS-K data as a Trend Study with adjustment for Start of School
2. Operationalizing Teacher Expectations as Intensity of Instruction and Proficiency Goal.
3. Attachment of student race and SES to the teacher model as modifiers of expectations.
4. Modeling of student learning without student race or SES
5. Learning and Forgetting functions in the Student Learning Curve model
6. Modeling a Teaching Curve

JEL Classifications:

C63 Computational Techniques • Simulation Modeling
I24 Education and Inequality
Z13 Social and Economic Stratification

PROBLEM STATEMENT 4

METHODS AND PROCEDURE 5

Data Source 5

The System Dynamics Modeling Process 7

PROBLEM ARTICULATION 9

Reference Modes 9

Exploratory Data Analysis 12

Opportunity for a Trend Study 12

Non-Linearity of Letter Recognition Probability Scores 15

The Problem Behavior 21

FORMULATION OF THE DYNAMIC HYPOTHESIS 22

Key variable and concepts 22

Operationalization of Teacher Expectations 22

Statement of the Dynamic Hypothesis 24

SPECIFICATION OF THE MODEL 24

Model Structure 24

Model Equations and Assumptions 26

Structure of Learning, Forgetting, and Teaching Curves 28

PARAMETER ESTIMATION 32

Model Initial Conditions 32

Estimation of Model Parameters by Model Reference Optimization 34

"Improving" the Fit for African American, SES5 41

DISCUSSION OF RESULTS 45

Model 'Validation' 45

Theil MSE Decomposition Comparison to Actual Data 47

Summary of 'Goodness of Fit" 50

An Unanticipated Validation 51

REFERENCE LIST 53

PROBLEM STATEMENT

The lines of research on Black/white reading gaps have divided into two: socio-cultural/out-of-school (poverty, parent's educational level, pre-natal care issues, childrearing and family support issues, African American culture, discrimination and stereotyping, etc.) and school-related/between-school/within-school (teacher expectations, teacher quality, rigor of curriculum and instruction, material resources, class size, student attendance, school diversity, etc.). There is acknowledgment that overlap exists between these two lines of research, but generally the focus is on either what the students bring to school with them or what they encounter when they get to school.

This study concentrates on the role that teacher expectations play in how teachers interact with their students. It does so with an explicit dynamic simulation methodology (System Dynamics or SD) aimed at exposing mechanisms leading to the accumulation of effects. And it does so via a dynamic simulation aided empirical analysis of data from a nationally representative sample.

The decision to focus on teacher expectation is the result of the use of a methodological approach we are unable to fully discuss here for lack of space (see Tauheed, 2013). The basic approach was to abstract from a possible model those factors that were deemed to be unlikely, given the circumstances generating the data, to have a differential effect on outcomes. For example, teacher quality, measured in level of professional certification or academic degrees, was deemed to be unlikely to have a differential effect on how well teachers were able to teach letter recognition to kindergarteners. The same logic was applied to other researched causal factors.

The decision to operationalize teacher expectations as variables representing teachers actions we see as consistent with the way teacher expectations was first revealed as a causal variable in what those expectations caused teachers to do, in the classroom, (Rist, 1970; Rosenthal, 1987; Rosenthal & Jacobson, 1968) as opposed to perceptions of teacher

bias or what teachers think, or say they think, about their students educational prospects (Ferguson, 2003).

METHODS AND PROCEDURE

Data Source

The data source for this study is the Early Childhood Longitudinal Study -Base Year (Kindergarten) database (ELCS-K), compiled by the National Center for Educational Statistics (NCES) of the U.S. Department of Education (National Center for Educational Statistics, 2001a).

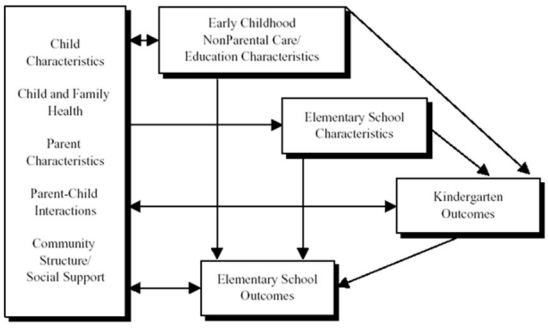
The ECLS-K data will be used for three primary purposes:

1. To supplement the development of "Reference Modes" for the pattern of achievement gaps for elementary school students by race.
2. To hypothesize structural relationships for System Dynamics modeling. The structural relationships will be hypothesized as a result of applying "Exploratory Data Analysis" techniques to the ECLS-K data, and by reference to educational research.
3. And, to estimate the values of parameters for the hypothesized structural relationships.

The ECLS-K data is a multistage probability sample of a cohort of students beginning in the kindergarten 1998-1999 school year, with the planned expansion of the dataset through the 5th grade. The sample contains data from approximately 22,000 students entering kindergarten in over 1000 schools nationwide. There is an over sample of Asian and Pacific Islander children and children in private schools.

Data are collected at the Child (child characteristics), Parent / Family (child and family health, parent-child interactions, parent characteristics, family characteristics), Community (community structure and social support), and School / Administration / Teacher (elementary school characteristics, elementary school outcomes, kindergarten outcomes, early childhood non-parental care/education characteristics) levels. The relationship between these categories is represented by the ECLS-K conceptual model (Figure 1).

Figure 1. ECLS-K conceptual model



Source: USER'S MANUAL FOR THE ECLS-K BASE YEAR PUBLIC-USE DATA FILES AND ELECTRONIC CODEBOOK - NCES 2001-029 (revised)

Children are assessed for physical, social, emotional and cognitive development by a number of instruments developed specifically for the ECLS-K, rather than pre-developed standardized tests. Parent/family, school/administration and teacher data are collected via structured interviews and questionnaires.

The performance scores used in this study derived from the Kindergarten Item Response Theory (IRT) scale scores in reading measured in the fall of 1998 and spring of 1999, and the probability scores for the cognitive skills comprising the IRT performance scores.

> IRT uses the pattern of right, wrong, and omitted responses to the items actually administered in a test and the difficulty, discriminating ability, and "guess-ability" of each item to place each child on a continuous ability scale. (National Center for Educational Statistics, 2001b, pp. 3-2)

> The IRT scale scores in the database represent estimates of the number of items students would have answered correctly if they had taken all of the 92 questions in the first- and second-stage reading forms (National Center for Educational Statistics, 2001b, pp. 3-4)

> Proficiency probability scores, although derived from the overall IRT model, are criterion-referenced measures of proficiency in specific skills. Because each proficiency score targets a particular narrow set of skills, they are ideal for studying the details of achievement, . . . They are useful as longitudinal measures of change because they show not only the extent of gains, but also where on the achievement scale the gains are taking place. **Thus, they can provide information on differences in skills being learned by different groups, as well as the relationships with processes, both in and out of school, that correlate with learning specific skills.** . . . Changes in proficiency probabilities over time may be used to identify the process variables that are effective in promoting achievement gains in specific skills. (National Center for Educational Statistics, 2001b, pp. 3-12, my emphasis)

The ECLS-K dataset used in this study is the Public rather than the Restricted use version. The differences between the two sets are to prevent disclosure of individual children, families, teachers, and/or schools, and consists of the following modifications (National Center for Educational Statistics, 2001b, pp. 1-6).

- Outlier values were top- or bottom- coded.
- Individual cases for which a particular variable posed an especially high risk of disclosure had the value of that variable altered (usually by no more than 5 to 10 percent) to reduce the risk;
- Some continuous variables were modified into categorical variables, and certain categorical variables had their categories collapsed; and
- Certain variables with too few cases and a sparse distribution were suppressed altogether, rather than modified.

Of the original 21260 child records, 3453 records of children who changed teachers or schools between the fall and spring testing rounds were excluded, leaving 17807 records. No other records were excluded from analysis, except those with missing values for relevant variables, explained below.

<u>The System Dynamics Modeling Process</u>

The process of developing a System Dynamics model of the research problem will follow standard procedure for the most part. This section describes this procedure and follows closely section 3.3 — Steps of the Modeling Process, from Sterman (2000).

Problem Articulation

Problem articulation will involve the definition of the problem as a set of 'Reference Modes' (Saeed, 1999) of key variables and concepts in order to set the boundary of the model within this study. The timeframe to be explored and the historical and expected future behavior of key variables will be defined.

Formulation of Dynamic Hypothesis

The formulation of the Dynamic Hypothesis will involve the development of a causal explanation of the dynamics of the problem as the endogenous consequence of the feedback structure.

Specification of the Simulation Model and Parameter Estimation

The formulation of a simulation model will involve the specification of structure and decision rules. It also involves the estimation of parameters and behavioral relationships for the model.

The estimation of parameters, behavioral relationships, and initial conditions will be accomplished using a number of techniques, including non-linear regression on actual data, and model reference optimization (Oliva, 2001) techniques which are non-linear programming techniques to perform parameter searches around an objective function.

Model Validation

Validation will involve the comparison of the developed model to Reference Modes, i.e. the expected or desired behavior of the model relative to the Dynamic Hypothesis. The model will also be compared to historical data to analyze "goodness of fit" and to suggest model improvements to improve fit. Comparison to actual data will be accomplished primarily using Theil statistics (Oliva, 1996; Sterman, 2000; Theil, 1966; Theil & Scholes, 1967).

PROBLEM ARTICULATION

Reference Modes

The historical behavior of the variables of interest may be captured by various methods, both quantitative and qualitative. The historical behavior of the variables of interest in SD models are typically described by using Reference Modes (Saeed, 1999). Reference modes are generally graphic descriptions, over time, of the behavior of the variables of interest and represent the general pattern and the interrelationships of the general patterns of these variables. Although Reference Modes represent these patterns over time, they are not to be confused with historical time series.

> "Indeed, a reference mode is a fabric of trends representing a complex pattern rather than a collection of historical time series. It may contain concrete variables actually existing in historical data as well as abstract variables summarizing qualitative information." (Saeed, 1999)

Saeed (1999, p. 3) defines four steps in developing Reference Modes from historical evidence which illustrates its abstract, qualitative and intuitive nature.

1. historical evidence used for defining a reference mode may consist of complex time paths with no discernable pattern or trend. These complex paths must be decomposed into simpler components to identify problematic patterns of behavior which may not necessarily be addressed in a single model. In this way, a model would address a problematic pattern and not merely replicate history.
2. a problem may often have multiple manifestations experienced over different geographic locations or over different periods of history. A penetrating problem description should collect these multiple manifestations and view them as multiple modes generated by the same system under different environmental conditions. A model capable of generating such multiple modes would often have in it the policy space to search for an effective means of change.
3. variables used in defining a reference mode may not exactly be the variables in a corresponding set of time series but may aggregate time series information in an abstract form. These variables must be delineated after carefully examining the decomposed components of historical data.

4. an end condition has often not been reached in the observed data when a problem is defined. In such cases, past trends must be projected intelligently into the future to be able to discern a complete pattern defined as a reference mode. A model subsuming both past and inferred future would have an appropriate time horizon for addressing the issue of change.

Prior ECLS-K Analyses

A prior analysis of the ECLS-K data was conducted by Reardon (2003), in what is a standard approach to analyzing and describing the achievement gap and its development.

Reardon performed a multilevel piecewise linear growth estimate of the Reading Scale Score African American-White gap from the ECLS-K, estimating both the initial kindergarten entry gap in Reading Scale Scores, and the slope of the learning rates, interpreted as points per month, for out-of-school, between-school, and within-school processes. These were estimated for the Pre-kindergarten, kindergarten and 1st grade periods and the intervening summer. Reardon's results are graphically summarized in Figure 2,

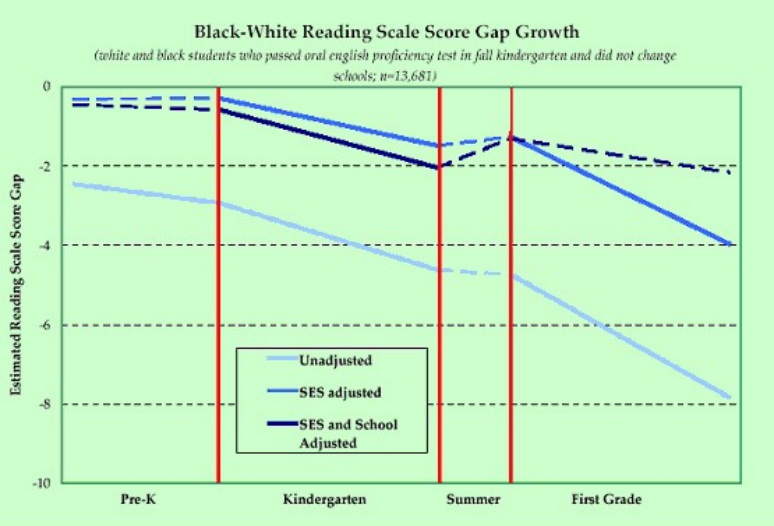


Figure 2. African American - White Reading Scale Score Gap Growth

Source: Sources of Educational Inequality: The Growth of Racial/Ethnic and Socioeconomic Test Score Gaps in Kindergarten and First Grade by Sear Reardon, 2003

Here we see the linear Reading Scale Score gap patterns for African American vs. White; unadjusted, adjusted for SES, and adjusted for SES and school effects. The White Reading Scale Score is represented by the horizontal axis at zero.

Reardon's estimates for the Reading Scale Score intercepts on entry to Kindergarten, (and the calculated percent differences), unadjusted, adjusted for Race, adjusted for SES and Race, and adjusted for SES, Race and School are shown in Table 1.

Table 1. African American - White Reading Scale Score Parameter Estimates

Parameter	Unadjusted	Race Adjusted	SES & Race Adjusted	SES, Race & School Adjusted
Intercept	19.867	20.919	19.882	19.577
African American		-2.932	-0.292	-0.578
Percent Difference		-14.02%	-1.47%	-2.95%

Reardon has noted that "one limitation of existing analyses using ECLS-K data is that they have relied exclusively on repeated cross-sectional analyses, an analytic strategy that does not take advantage of the longitudinal strength of the data" (Reardon, 2003, p. 5, emphasis added). Reardon's use of the term "repeated cross-sectional analyses", referring to the fall/spring repeated waves of assessments, and the use in this study will be discussed below. Reardon's linear growth model still misses the opportunity to utilize the repeated (weekly) cross-sectional nature of the research design to investigate the shape of the learning curves.

As estimated by Reardon, on entry to kindergarten, African American students score approximately 2.9 points (unadjusted, $p<.001$) lower than Whites on the reading assessment. African American children at entry to kindergarten have the same Reading Scale Scores as Whites when adjusted for SES. This result, again consistent with other ECLS-K studies (Fryer & Levitt, 2002; Lee et al., 2002), is important in this study for specification of the SD model to be simulated. By the end of kindergarten, the unadjusted reading gap has grown to 4.6 points.

There is no significant summer change in the unadjusted reading gap estimated at 4.7 points on entry to 1st grade. Other analyses of ECLS-K data (Fryer & Levitt, 2002; Lee et al., 2002) show a similar pattern of scale score gap growth.

Exploratory Data Analysis

Opportunity for a Trend Study

However, while we see linear growth over the entire sample, there is no theoretical reason to assume linear scale score growth for any particular sub-sample. Had the ECLS-K study been designed explicitly to study the shape of scale score growth for individual students, there would have been repeated assessments of each student during the semester. This was not done. However, because of the way the ECLS-K study was conducted there is an opportunity to gain insight into the shape of scale score growth using fairly short time periods (weekly) during the fall and spring Kindergarten semesters, when scores are aggregated by groups of interest (racial and socio-economic), and if we make the assumption that the research design and data satisfy the requirements of a "repeated cross-sectional study", also known as a "trend study" (Taris, 2000, p. 6) where "two or more cross-sectional studies are conducted at two or more occasions" (2000, pp. 6-7).

The reference to "repeated cross-sectional analysis" in the previous quote by Reardon is not in the same context as the trend study referred to by Taris. Reardon is referring to previous ECLS-K studies that have taken the fall kindergarten and spring kindergarten scores as two cross-sectional periods, one each term. Thus, the assessment times, and student length of time in school are not taken into account in the analysis therefore "not [taking] advantage of the longitudinal strength of the data" (Reardon, 2003, p. 6). Taris' use of the term repeated cross-sectional/trend study takes each assessment period as a separate cross-sectional sample and "thus, the trend study is . . . cross-sectional at the level of the sample units, but longitudinal at the level of the research units" (Taris, 2000, p. 6).

Requirements of a Trend Study

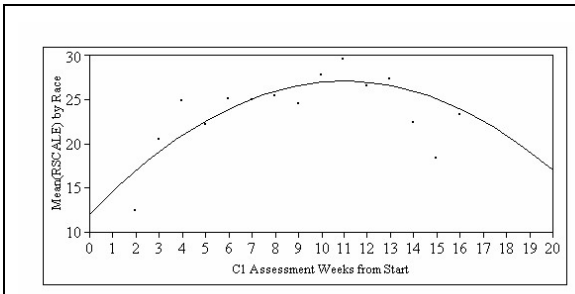
A requirement for the validity of a Trend Study is that the cross-sectional participants be compatible in terms of relevant characteristics, this requirement is met in this study by analysis of score trends by race/ethic group and class (SES). Another requirement is that the same measurement instrument be used throughout. The ECLS-K study meets this requirement.

Another threat to validity is the concern that students be assigned randomly to each cross-sectional period. The ECLS-K study also meets this requirement. Students, actually schools, were assigned to testing dates as part of the recruitment process into the study.

> Once the school administrator agreed to participate, he or she was asked to set an appointment for two visits by the ECLS-K field staff to the school in the fall of the 1998-99 school year. The first visit, the preassessment visit, was to select the sample of children . . . the second visit was to conduct the child assessments. (National Center for Educational Statistics, 2001a, pp. 5-3)

There is no reason to believe that this assignment method would assign any relevent group of students to a particular testing date, and to thereby bias the scale score growth pattern with regards to any grouping, including race or socio-economic groupings. That being the case we would expect each group of students tested as the semester proceeds to represent a random assignment from the sample of students, and for the distribution of scores of such to display the estimated growth path of students on average.

Plotting fall Reading Scale Scores by Assessment Weeks from Start for Asian, African American, Hispanic and White students generates Figure 5 through Figure 8.



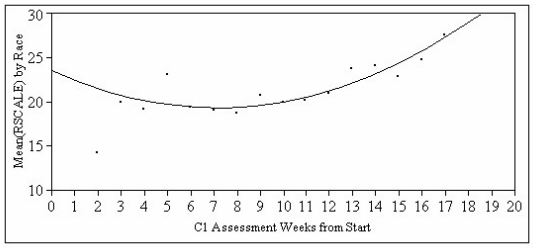


Figure 5. Fall Reading Scores by Assessment Weeks from Start (Asian)

Figure 6. Fall Reading Scores by Assessment Weeks from Start (African American)

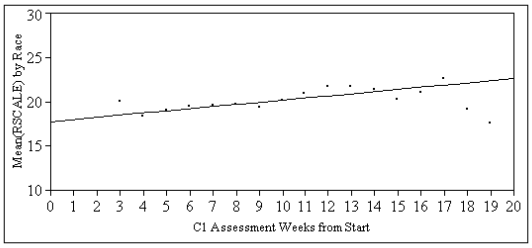


Figure 7. Fall Reading Scores by Assessment Weeks from Start (Hispanic)

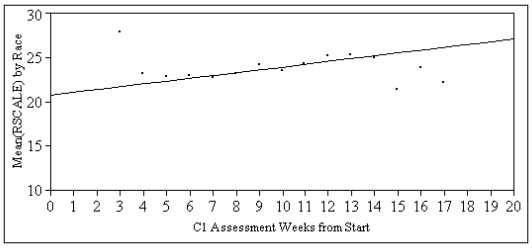


Figure 8. Fall Reading Scores by Assessment Weeks from Start (White)

The fits are in Table 7 through Table 10. The standard errors and t scores were estimated from the individual scores not the means. Notably the fits for Asian and African American students are decidedly non-linear with Asian students showing a statistically significant negative quadratic term reflecting an eventual slow-down and decrease in proficiency. African American students show a statistically significant negative first-order term. This would mean that African American students show a decline in performance before showing an increase.

Table 7. Fit of Fall Reading Scores by Assessment Weeks from Start (Asian)

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	12.0860	4.3279	2.7900	0.0053
C1 Assessment Weeks from Start	2.7288	0.9837	2.7700	0.0057
C1 Assessment Weeks from Start^2	-0.1237	0.0537	-2.3000	0.0216

Table 8. Fit of Fall Reading Scores by Assessment Weeks from Start (African American)

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	23.5926	1.9425	12.1500	<.0001
C1 Assessment Weeks from Start	-1.1661	0.4235	-2.7500	0.0059
C1 Assessment Weeks from Start^2	0.0814	0.0226	3.6100	0.0003

Table 9. Fit of Fall Reading Scores by Assessment Weeks from Start (Hispanic)

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	17.8793	0.6377	28.0400	<.0001
C1 Assessment Weeks from Start	0.2406	0.0690	3.4900	0.0005

Table 10. Fit of Fall Reading Scores by Assessment Weeks from Start (White)

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	20.8439	0.3598	57.9300	0.0000
C1 Assessment Weeks from Start	0.3156	0.0387	8.1500	<.0001

Non-Linearity of Letter Recognition Probability Scores

Plotting the fall term average Reading Scale Scores by race reveal non-linear time paths. A further analysis investigates the time trend for the fall and spring terms and for grouping by SES as well as race. We investigate this trend not for the Reading Scale Scores, but for the components of the scale score, the Proficiency probability scores, which as previously stated "are ideal for studying the details of achievement, rather than the single summary measure provided by the scale scores." And such scores "are useful as longitudinal measures of change because they show not only the extent of gains, but also where on the achievement scale the gains are taking place." "Thus, they can provide information on differences in skills being learned by different groups." (National Center for Educational Statistics, 2001b, pp. 3-12).

The proficiency score we analyze and model in this study is the Letter Recognition Probability Score, which is the first cognitive skill measure comprising the Reading Scale Score. The subsequent proficiency scores, Beginning Sounds, Ending Sounds, Sight Words, and Words in Context, are not modeled in this study, and remain subjects for further research. We also analyze and directly model African American and White scores, proposing a method for extending the model to handle the differences between those scores and Asian and Hispanic scores, in further research.

We aggregate the Letter Recognition Probability Scores (RPROB1L) by Assessment Week from Start, Race and SES. The means and standard deviations for the African American RPROB1L are shown in Table 11, those for White in Table 12.

Table 11. Mean and Standard Deviation of Letter Recognition Probabilities African American by SES

	Mean					Standard Deviation				
AWS	SES1	SES2	SES3	SES4	SES5	SES1	SES2	SES3	SES4	SES5
5	0.1881	0.5008	0.5627	0.8727	0.8561	0.3568	0.5765	0.4954	0.2989	0.3169
6	0.3873	0.5659	0.7376	0.6833	0.7845	0.4184	0.4336	0.3688	0.3859	0.3608
7	0.2963	0.5897	0.6093	0.6803	0.6903	0.4013	0.4240	0.4058	0.4164	0.3856
8	0.3435	0.4192	0.5556	0.6943	0.6646	0.4197	0.3965	0.4331	0.4001	0.4030
9	0.4408	0.5843	0.6864	0.7588	0.8444	0.4230	0.4247	0.4122	0.3611	0.3314
10	0.4421	0.5063	0.6237	0.7760	0.7500	0.4267	0.4363	0.4388	0.3708	0.3981
11	0.4241	0.5663	0.7296	0.7065	0.7810	0.4221	0.4139	0.3923	0.4032	0.3730
12	0.5183	0.7084	0.7287	0.8024	0.8014	0.4425	0.3735	0.4018	0.3725	0.3607
13	0.5827	0.9166	0.8326	0.7684	0.9364	0.4366	0.2103	0.3481	0.3664	0.1142
14	0.6390	0.8226	0.7742	0.6860	1.0000	0.5530	0.3193	0.3832	0.4170	0.0000
30	0.9975	0.9868	0.9863	0.8625	1.0000	0.0007	0.0258	0.0228	0.1945	NA
31	0.6748	0.9365	0.8945	0.9411	0.9996	0.4466	0.0937	0.3102	0.1856	0.0005
32	0.7633	0.8805	0.9083	0.9074	1.0000	0.3891	0.2535	0.2418	0.2849	0.0000
33	0.8655	0.8430	0.9914	0.8826	0.9955	0.2972	0.3291	0.0184	0.3022	0.0200
34	0.8042	0.8874	0.9237	0.9447	0.8987	0.3393	0.2792	0.2235	0.1826	0.2825
35	0.9096	0.9000	0.9296	0.9371	0.9645	0.2232	0.2532	0.2186	0.2059	0.1483
36	0.8739	0.9007	0.9455	0.9578	0.9564	0.2616	0.2543	0.1866	0.1646	0.1753
37	0.8562	0.9306	0.9465	0.9832	0.9099	0.3077	0.2146	0.2016	0.0574	0.2624
38	0.8296	0.8960	0.9584	0.9754	0.9830	0.3256	0.2795	0.1473	0.1025	0.0380
39	0.8094	0.9187	0.9080	0.9407	0.9999	0.3509	0.2491	0.2625	0.2081	0.0003
40	0.6119	0.8920	0.9156	0.9489	0.9969	0.4410	0.3135	0.1876	0.1391	0.0096

Table 12. Mean and Standard Deviation of Letter Recognition Probabilities White by SES

	Mean					Standard Deviation				
AWS	SES1	SES2	SES3	SES4	SES5	SES1	SES2	SES3	SES4	SES5
5	0.2884	0.4369	0.6112	0.7234	0.8181	0.4089	0.4506	0.4138	0.3814	0.3464
6	0.3740	0.5856	0.6501	0.7345	0.8490	0.4315	0.4250	0.4159	0.3835	0.3076
7	0.4785	0.5622	0.6791	0.7135	0.8707	0.4340	0.4346	0.3975	0.3861	0.2830
8	0.4811	0.5715	0.6292	0.8076	0.8428	0.4359	0.4276	0.4196	0.3379	0.3126
9	0.4253	0.5607	0.6991	0.7938	0.8932	0.4410	0.4399	0.4041	0.3500	0.2539
10	0.4654	0.6490	0.6520	0.7686	0.8588	0.4429	0.4125	0.4186	0.3630	0.3054
11	0.4638	0.6704	0.7516	0.8079	0.8876	0.4357	0.4062	0.3687	0.3421	0.2767
12	0.5849	0.6962	0.7577	0.8323	0.8761	0.4326	0.4069	0.3764	0.3173	0.2803
13	0.5884	0.7098	0.8130	0.8310	0.8957	0.4541	0.4007	0.3372	0.3159	0.2545
14	0.4004	0.7517	0.8505	0.8586	0.9823	0.4242	0.3600	0.3086	0.2752	0.0360
30	0.6928	0.9692	0.9642	0.9678	0.9885	0.4444	0.1526	0.1755	0.1469	0.0960
31	0.9333	0.9225	0.9136	0.9647	0.9814	0.1103	0.2188	0.2427	0.1368	0.1137
32	0.8262	0.9219	0.9506	0.9729	0.9761	0.3390	0.2522	0.1843	0.1367	0.1261
33	0.8136	0.9112	0.9565	0.9764	0.9810	0.3397	0.2322	0.1720	0.1284	0.1075
34	0.8757	0.9264	0.9568	0.9716	0.9895	0.3006	0.2243	0.1709	0.1394	0.0880
35	0.8848	0.9344	0.9609	0.9779	0.9900	0.2836	0.2168	0.1576	0.1193	0.0699
36	0.8542	0.9311	0.9574	0.9737	0.9860	0.3167	0.2165	0.1770	0.1389	0.0860
37	0.8613	0.9449	0.9450	0.9755	0.9898	0.3079	0.1928	0.1825	0.1251	0.0746
38	0.9057	0.9597	0.9512	0.9804	0.9777	0.2510	0.1635	0.1712	0.1160	0.1333
39	0.8642	0.9223	0.9714	0.9790	0.9974	0.3176	0.2251	0.1440	0.1162	0.0154
40	0.8529	0.8413	0.9740	0.9800	0.9999	0.2803	0.3473	0.1137	0.1164	0.0007

Non-Linear Exploration of Time Trends from Actual Data

Given the possible non-linear nature of the spring through fall term Reading Scale Scores and given the natural constraint placed on the proficiency probability scores (in range 0-1.0), an explicitly non-linear model is used to explore the time trend. The model used here is a variation of a Gaussian Amplitude curve. The equation for the standard Gaussian Amplitude curve is Equation 1, plotted in Figure 9. (Tobias, 2005)

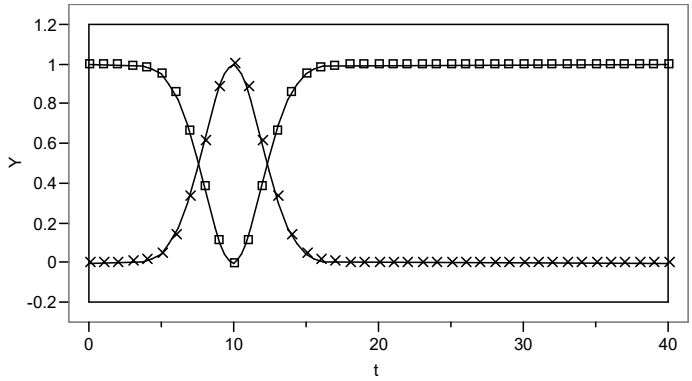
In our reformulation, Equation 1 for RPROB1L takes the form of Equation 2, where the curve is inverted by taking A ≡ Ymin - Ymax, making the asymptote y0 ≡ Ymax. Xmin represents the X position of Ymin, which if negative, fits a monotonically ascending curve for t = Assessment Weeks from Start > 0. W is a slope related parameter of the width of the curve at ½ amplitude. Equation 2 is also plotted in Figure 9.

Equation 1. Gaussian Amplitude Function

$$y = y_0 + Ae^{\frac{-(X-X_c)^2}{2*W^2}}$$

Equation 2. Inverse Gaussian Amplitude Function

$$Y(t) = RPROB1L(t) = Y_{max} + (Y_{min} - Y_{max}) * e^{\frac{-(t-X_{min})^2}{2*W^2}}$$



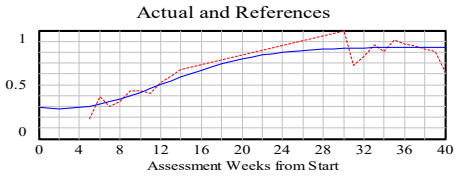
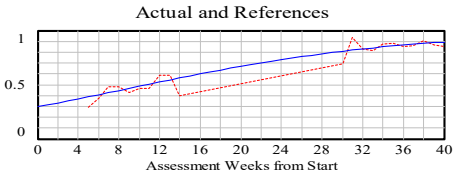
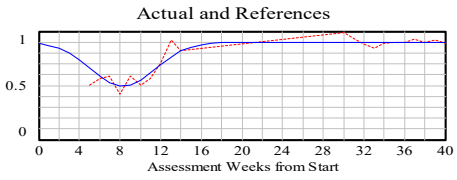
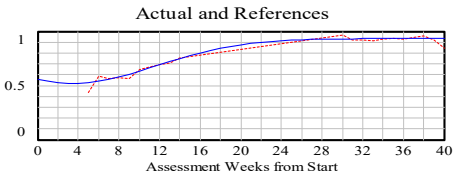
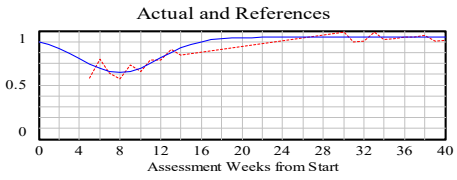
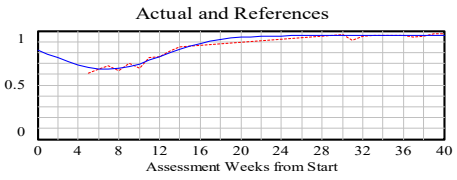
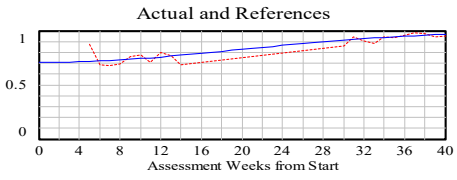
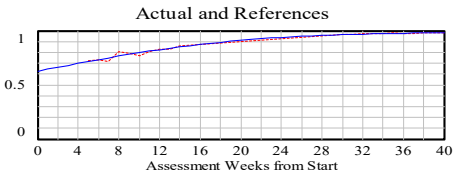
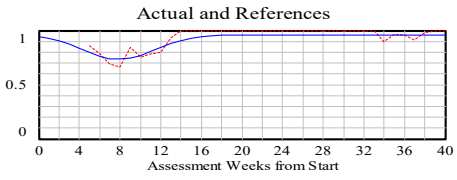
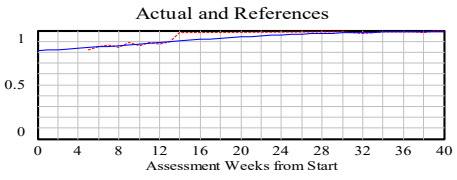
× = Gaussian Amplitude Curve - □ = Inverse Gaussian Amplitude Curve

Figure 9. Gaussian and Inverse Gaussian Amplitude Curves

This curve as modified was chosen because of its ability to fit asymptotically to Ymax, to fit the apparently declining/ascending portion of the African American SES2, SES3, and possibly SES5 curves, but to also fit the continuously ascending other curves by estimating negative Xmin. However, as we will see, a disadvantage of this curve is the symmetric nature of the fit of the slope on the left- and right-hand sides of Xmin. Since there is only one slope related parameter (W), when the fit is influenced heavily by the right hand side data, the left-hand side slope is forced to mirror the right-hand side slope which distorts the left hand fit. Because of this and because data points before week five are non-existent, we cannot say anything with confidence about the estimated intercept, and although we can calculate one, do not generally do so. The intercept is to be estimated by simulation of the System Dynamics model.

The results of the Inverse Gaussian fits for African American and white students are plotted in Figure 10. The fit statistics (parameter estimate, approximate standard error, t statistic, and t probability) are in Table 13.

Figure 10 - Actual Data and Reference Modes for RPROB1L by Race

SES	Actual Data and Reference Modes for African American	Actual Data and Reference Modes for White
1		
2		
3		
4		
5		

Since we are in the exploratory phase the purpose of fit statistics is not to determine statistical significance but to ascertain possible nonlinearity beyond visual inspection. Inspection of Figure 10 tells us that some curves show non-linearity and some do not.

The fits in Table 13 allow us to have greater confidence in our visual inspections.

Table 13 - Inverse Gaussian Fit of RPROB1L by Race

SES	Inverse Gaussian Fit of RPROB1L for African American					Inverse Gaussian Fit of RPROB1L for White				
1	Parameter	Estimate	Approx Std Err	t score	Prob>t	Parameter	Estimate	pprox Std I	t-score	Prob>t
	Ymin	0.2788	0.1866	1.4943	0.1353	Ymin	0.1441	1.0379	0.1388	0.8896
	Ymax	0.8461	0.0171	49.4796	0	Ymax	0.9652	0.2353	4.1016	0
	Xmin	2.2709	9.6639	0.235	0.8142	Xmin	-16.4713	62.1881	-0.2649	0.7912
2	Parameter	Estimate	Approx Std Err	t score	Prob>t	Parameter	Estimate	pprox Std I	t-score	Prob>t
	Ymin	0.495	0.0224	22.1275	0	Ymin	0.5197	0.0616	8.4321	0
	Ymax	0.9	0.0148	60.9933	0	Ymax	0.9334	0.0084	110.8881	0
	Xmin	8.307	0.3388	24.5164	0	Xmin	3.471	4.1295	0.8405	0.4007
3	Parameter	Estimate	Approx Std Err	t score	Prob>t	Parameter	Estimate	pprox Std I	t-score	Prob>t
	Ymin	0.6148	0.0212	29.0209	0	Ymin	0.6457	0.0102	63.333	0
	Ymax	0.9406	0.0148	63.4048	0	Ymax	0.9556	0.0069	138.0928	0
	Xmin	7.9244	0.6816	11.6263	0	Xmin	6.8436	0.7502	9.1223	0
4	Parameter	Estimate	Approx Std Err	t score	Prob>t	Parameter	Estimate	pprox Std I	t-score	Prob>t
	Ymin	0.7096	0.0959	7.4003	0	Ymin	0.4207	1.2423	0.3387	0.7349
	Ymax	1	0			Ymax	0.9866	0.0291	33.9562	0
	Xmin	0.6546	17.633	0.0371	0.9704	Xmin	-18.0002	59.3093	-0.3035	0.7615
5	Parameter	Estimate	Approx Std Err	t score	Prob>t	Parameter	Estimate	pprox Std I	t-score	Prob>t
	Ymin	0.7347	0.029	25.3024	0	Ymin	0.7962	0.1511	5.2692	0
	Ymax	0.959	0.0196	48.9541	0	Ymax	0.9959	0.0211	47.1419	0
	Xmin	7.8883	0.9259	8.5195	0	Xmin	-7.172	26.114	-0.2746	0.7836

There are similarities and differences between the curves to be noted. It is helpful to note these similarities and differences with regards to the time periods in which they were generated. Time period #1 (TP1) from week 5 through week 14) and time period #2 (TP2) from week 30 through week 40). We note that:

1. All Reference Modes are level (or nearly so) by week 30 of TP2, with the exception of African American SES4, and White SES1. The leveling of the Letter Recognition proficiency score is expected to occur eventually, considering the upper limit on probabilities. However, all curves reach a

maximum in order by SES, with White values systematically higher than African American values - (0.8461, 0.9000, 0.9406, visually .9590, and 0.9590 for African American, and visually 0.9, 0.9334, 0.9556, 0.9866, and 0.9959 for White). This occurs even though students are still in school after week 30 and have 10 weeks of time left in TP2 to improve in proficiency.

2. African American SES2, SES3 and SES5 show statistically significant positive Ymin and Xmin values estimated within the range of TP1, while White SES3 Ymin and Xmin values are also statistically significant and positive within TP1. This denotes a decline in proficiency before an improvement.
3. The curves for African American SES1 and SES4, and White SES1, SES2, SES4, and SES5 are decidedly linear within the range of TP1. The curves for African American SES2, and SES3, and White SES3 are decidedly non-linear within the range of TP1.

The Problem Behavior

While prior research on the African American/White reading achievement gap using the ECLS-K data has reported a gap in the initial reading performance of African American vs. White kindergarten students, and have reported an increase in the gap during the kindergarten year, they have estimated this gap using some version of linear regression modeling, controlling for various explanatory factors such as SES to explain individual variation.

However, by imposing linear models, and looking for an explanation of variation in individual differences, this research has missed the opportunity to study the trend in the data of RSSs for African American and Whites student, and for students of different SES categories, as the kindergarten school year progresses.

Due to the previously noted disadvantages of the non-linear model chosen, especially as regards the single slope, the reference modes generated by regression represent an idealization of the actual performance trends for students by race and SES. They however are an indication that assuming the same growth path for all subgroups, linear or otherwise, is empirically unjustified. The display of the RSS trends demonstrates the decidedly non-linear growth trends of these scores for African American vs. White students.

The problem that this study is confronted with is how to systematically model the non-linear trends in the Letter Recognition Probability Scores of these RSSs, as shown in the Reference Modes of Figure 11. As a system, the same dynamics must drive the generation of all 10 Reference Modes. There can be no special cases to be explained by a different set of dynamics.

Systematically explaining the curve characteristics is the problem to be modeled.

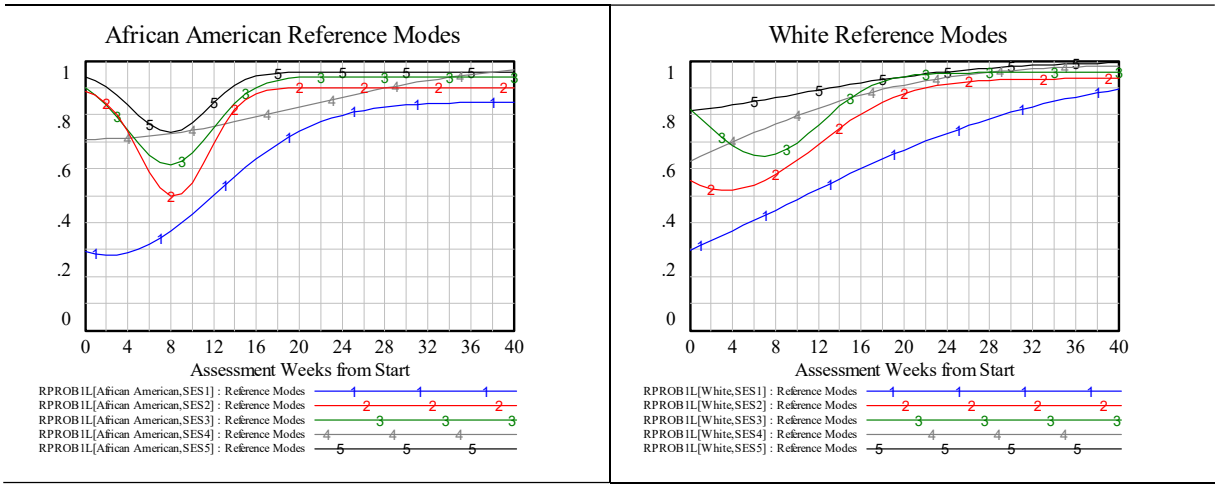


Figure 11. Reference Modes for RPROB1 by Race and SES

FORMULATION OF THE DYNAMIC HYPOTHESIS

Key variable and concepts

The effect of teacher expectations varied by student race and socioeconomic status (SES) and its effect on student learning is the key hypothesized causal relationship to be modeled.

Operationalization of Teacher Expectations

Student's SES (the SES of the student's parents), is in two forms, a quintile categorization of parent's education, occupation, and income and the continuous SES measure from which quintiles are calculated. Prior to entering kindergarten, the parent's SES category is an indication of the existence of resources which can be used by the parent

to affect the child's level of preparation for school, and our model will reflect this as the initial level of learning the child has attained on entry to kindergarten.

It is hypothesized that once the child enters kindergarten, the student's SES category becomes a signal triggering a set of precepts concerning the ability of students of each SES category to benefit from instruction. Thus, for the teacher, the student's SES becomes a variable, affecting the teacher's expectations for student learning in the educational process.

Within a classroom this can result in all students receiving the same Intensity of Instruction (measured in new letters of the alphabet to be learned per week) and Proficiency Goal (expected final proficiency probability) based on class average (mean, median or mode) race and SES composition; or with teachers separating students into groupings, based on expected performance 'biased' by student race and SES (Rist, 1970), or by teachers dividing their time among students biased by the race and SES of the student (Anyon, 1980; Ferguson, 2003; Jussim & Harber, 2005; Rist, 1970). As indicated by Table 14, calculated from ECLS-K data, African American students are in classrooms where 68% of students are African American, while White students are in classrooms where 84% of students are White.

Table 14. Classroom Cross Attendance of Students by Race

Race	N	% Asian	% African American	% Hispanic	% White
Asian	1201	0.4594	0.0707	0.1594	0.3105
African American	2533	0.0335	0.6783	0.0815	0.2066
Hispanic	3144	0.0609	0.0657	0.6387	0.2347
White	9906	0.0376	0.0528	0.0745	0.8350

Table 15 indicates that SES1 students are in classrooms where 52% of students are SES1, while SES5 students are in classrooms where 52% of students are also SES5. The percentages for SES2-SES4 are lower but still demonstrate homogeneity of classrooms by SES.

Table 15. Classroom Cross Attendance of Students by SES

SES	N	%SES1	%SES2	%SES3	%SES4	%SES5

%SES1	3050	0.5180	0.1960	0.1320	0.0961	0.0580
%SES2	3285	0.1819	0.3848	0.1850	0.1447	0.1035
%SES3	3386	0.1189	0.1795	0.3626	0.1888	0.1503
%SES4	3547	0.0826	0.1340	0.1802	0.3765	0.2267
%SES5	3812	0.0464	0.0892	0.1335	0.2109	0.5200

Because students in this study attend school in classrooms primarily with students of their own race and SES, either method (averaging or grouping) for 'assignment' of Intensity and Goals to students results in substantially the same instruction pattern based on student race and SES.

This study operationalizes teacher expectations of students by race and SES as the application of differential Intensity of Instruction and Proficiency Goals affecting teacher Level of Instruction.

Statement of the Dynamic Hypothesis

Systematic variation of Intensity of Instruction and Proficiency Goal as a function of student's race and SES, biases teacher's expectations, systematically affecting student learning, resulting in the previously described problematic curve characteristics shown in reference modes of Letter Recognition Probability Scores differentiated by race and SES.

SPECIFICATION OF THE MODEL

Model Structure

The model diagram in Figure 12 (modeled using Vensim® System Dynamics software) and specified in Equation 3 thru Equation 7 models a teacher and student interacting within a classroom. We model flow (rate) variables: Student Rate of Learning [RL(t)], Student Rate of Forgetting [RF(t)], and Teacher Rate of Instruction [RI(t)], and stock (level) variables: Student Level of Learning/Letter Recognition Probability/Proficiency [RPROB1L(t)] and Teacher Level of Instruction [RPROB1I(t)].

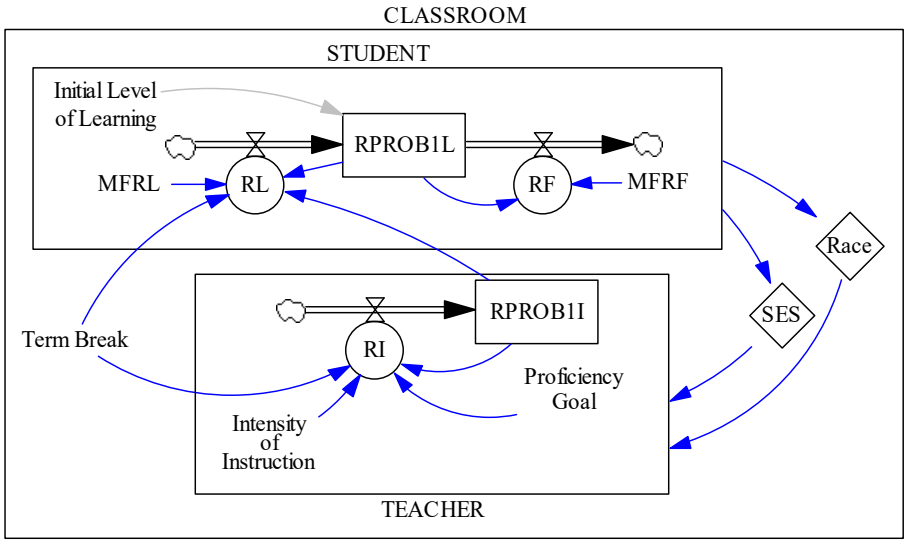


Figure 12. Model of Teacher/Student Letter Recognition Learning Transaction

The student learning process is represented in the model by RPROB1L(t), RL(t), RF(t), and Initial Level of Learning.

The teacher instruction process is represented by RPROB1I(t), RI(t), and Intensity of Instruction, and Proficiency Goal.

The 'outflow' from the teacher to the student is instruction in letter recognition measured by the Level of Instruction (RPROB1I). The 'inflow' from the student to the teacher is the race and SES of the student, which moderates the teacher's Intensity of Instruction, and Proficiency Goals.

Term Break (discussed below) interrupts the effect of RI on RPROB1I, and the effect of RL on RPROB1L.

Maximum Fractional Rate of Learning (MFRL) and Maximum Fractional Rate of Forgetting (MFRF) (discussed below) are rate variation parameters for RL and RF.

Model Equations and Assumptions

In addition to the functional forms of the learning, forgetting and teaching curves (discussed below), the model makes theoretical assumptions modeled in the subscript structure of the model equations.

Equation 3. Student Level of Learning

$$RPROB1L(t)_{[Race,SES]} = \int_{t=0}^{40} (RL(t)_{[Race,SES]} - RF(t)_{[Race,SES]})\ dt$$

$$where\ RPROB1L(0)_{[Race,SES]} = Initial\ Level\ of\ Learning_{[SES]}$$

The Initial Level of Learning is subscripted by SES only, modeling the assumption that on entry to kindergarten, children of different races, but the same SES have the same Initial Level of Learning on average. The adoption of this assumption is justified by SES adjusted RSSs estimated by previous ECLS-K studies (Fryer & Levitt, 2002; Lee et al., 2002; Reardon, 2003)

Equation 4. Student Rate of Learning

$$RL(t)_{[Race,SES]} = MFRL * (RPROB1I(t)_{[Race,SES]} - RPROB1L(t)_{[Race,SES]})$$
$$if\ \neg Term\ Break\ and\ (RPROB1I(t)_{[Race,SES]} > RPROB1L(t)_{[Race,SES]});$$
$$otherwise\ 0$$

The model assumes that until the teacher's Level of Instruction has reached the current Student Level of Learning/Letter Recognition Probability/Proficiency, the teacher's instruction is ineffective in raising the student's level of proficiency since no new letters are taught/learned. Equation 4 - where $RL(t)_{[R,S]} = 0$ until $(RPROB1I(t)_{[R,S]} > RPROB1L(t)_{[R,S]})$, models this assumption.

Equation 5. Student Rate of Forgetting

$$RF(t)_{[Race,SES]} = \frac{MFRF * RPROB1L(t)_{[Race,SES]} * (1 - (RPROB1L(t)_{[Race,SES]})^{(1.1-1)})}{(1.1-1)}$$

The Maximum Fractional Rate of Learning (MFRL) and the Maximum Fractional Rate of Forgetting (MFRF) are un-subscripted, modeling the assumption that on average they are the same for all students, regardless of race and SES.

Equation 6. Teacher Level of Instruction

$$RPROB1I(t)_{[Race,SES]} = \int_{t=0}^{40} RI(t)_{[Race,SES]}\ dt\ ; \quad where\ RPROB1I(0)_{[Race,SES]} = 0$$

The teacher's Level of Instruction starts at zero. The functional form choice is discussed below.

Equation 7. Teacher Rate of Instruction

$$RI(t)_{[Race,SES]} =$$
$$Intensity\ of\ Instruction_{[Race,SES]} * (Proficiency\ Goal_{[Race,SES]} - RPROB1I(t)_{[Race,SES]})$$
$$if\ \neg Term\ Break;$$
$$otherwise\ 0$$

The average Intensity of Instruction, and the average expected teacher proficiency goal for student performance (Proficiency Goal) are subscripted by **student's** race and SES.

Additional Model Assumptions

The Term Break represents the three-week Christmas/New Year's break between the Fall and Spring terms. This break comes after the 15th week of school for all students. The placement was based on taking the number of weeks from the median for School Start Dates (August 31) and counting the number of completed weeks (from 0) until the first week before Christmas. The break spans from that week until one week after New Year's Day. While the placement of this break is not accurate for all students, the most significant impact of the break is to introduce a theoretically necessary discontinuity in instruction and learning into the model, after all of the Fall term data points (the 5th through the 14th week) have been used in the estimation process.

Structure of Learning, Forgetting, and Teaching Curves

Learning Curves

Learning curves, the relationship of instruction time to learning level, are a widely researched area in Cognitive Psychology, Industrial Engineering, and Management Science. Yet, learning curves have little appearance in the education literature, except with reference to teacher training and productivity. Two early studies on the learning curves of young children are Virginia Nelson's An Analytical Study of Child Learning (1936), and William Beasley's Visual Pursuit in 109 White and 142 Negro Newborn Infants (1933). 'Recent' work in this area has been done by economists Byron Brown and Daniel Saks: Measuring the Effects of Instructional Time on Student Learning: Evidence from the Beginning Teacher Evaluation Study and The Microeconomics of Schooling (1981, ; 1986). Learning curves enter the economics literature in the work of Kenneth Arrow (1962) and Nicholas Balof (1966).

The learning curve models used in this study come from the work of Jaap Murre and Antonio Chessa. Murre and Chessa's model of simple or non-interactive learning (Murre & Chessa, submitted-b) defines the probability of recall at time t after a learning trial as, $p(t) = 1 - e^{-vl\tilde{r}(t)}$, "where v is the learning rate parameter, l is the learning time, and $\tilde{r}(t)$ expresses the decline of intensity with time t since the end of the learning trial" (submitted-b, 8). In their model of learning and forgetting, forgetting is represented by the $\tilde{r}(t)$ function, which they assume to be constant if learning trials occur at fixed intervals, and therefore in their model the "effect [of $\tilde{r}(t)$] is subsumed under v" (submitted-b, 8)

yielding $p(t) = 1 - e^{-vl}$, making the v parameter a combined learning rate/forgetting rate parameter. In our model we maintain the learning rate property of v, defining the forgetting property separately. Thus, using the revised Murre and Chessa model of non-interactive learning we have the recall probability without forgetting at time t after a learning trial as $p(t) = 1 - e^{-vt}$. The recall probability at time t is equal to the proficiency probability at the time of an assessment at time t. We therefore restate this as $RPROB1L(t) = 1 - e^{-vt}$. Since $RPROB1L(t) - 1 = -e^{-vt}$, the rate is

$RL(t) = dRPROB1L/dt = ve^{-vt} = v(1 - RPROB1L(t))$.

In Murre and Chessa's model of non-interactive learning, a learning trial involves teaching the subject to the point of saturation, i.e. where p(0) = 1.0, representing 100% recall, which will be reached if the learning trial is long enough. In our model the learning trial for RPROB1L consists of the entire sequence of instruction in new letters of the alphabet (along with review of prior letters). The learning trial produces increasing levels of RPROB1L as the learning trial proceeds. Thus, the Level of Instruction at time t, RPROB1I(t), becomes the upper limit to the level of RPROB1L(t), which the student approaches, but may not reach, rather than the upper limit of 100% proficiency in Murre and Chessa's learning trial. **With the learning rate parameter v, equal to MFRL**, this reformulates our learning rate function as $RL(t) = MFRL\ (RPROB1I(t) - RPROB1L(t))$, which we see in Equation 4. This function defines a positive feedback between the student's Rate of Learning(RL(t)) and the teacher's Level of Instruction (RPROB1I(t)), but a negative feedback between the student's Rate of Learning and the student's Level of Learning proficiency (RPROB1L(t)).

Forgetting Curves

Murre and Chessa's (submitted-a) model of forgetting defines the probability of recall at time t after a learning trial as, $p(t) = 1 - e^{-r(t)}$ (submitted-a, 17), "where $r(t) = \mu\tilde{r}(t)$" (submitted-a, 12). The parameter μ represents the physical intensity of memory representations in Murre and Chessa's Poisson point process model of memory retention. The function $\tilde{r}(t)$ is as stated above, where it "expresses the decline of intensity with time t since the end of the learning trial" (Murre & Chessa, submitted-b, 18). The functional form for $\tilde{r}(t)$ is "an exponential decline function . . . $\tilde{r}(t) = e^{-at}, a > 0$" (23). Therefore, the probability recall function takes the form $p(t) = 1 - e^{-\mu e^{-at}}$. The function $e^{-\mu e^{-at}}$ is the functional form for a Gompertz curve, which is in the Richards family of curves, and is modeled as a modified Richards curve with Gompertz parameters (see Sterman, 2000, 299). The rate function for the Richards curve as modeled by Sterman, takes the form $\frac{dP}{dt} = \frac{gP}{(m-1)}\left[1 - \left(\frac{P}{C}\right)^{m-1}\right]$, where C is the capacity parameter, and g is the maximum rate parameter. Since the C parameter represents the capacity (maximum asymptote) of the function and is 1 (100% probability) in our model, and this function approaches the Gompertz curve as m approaches one[2], our version of this function defines the Rate of Forgetting as $RF(t) = \frac{dP}{dt} = \frac{g\,P\left[1 - (P)^{1.1-1}\right]}{(1.1-1)}$. **Which is the function in Equation 5, with g = MFRF**, and P(t)=RPROB1L(t). This function defines positive feedback between the student's Rate of Forgetting (RF(t)) and the student's Level of Learning proficiency (RPROB1L(t)), and negative feedback between the student's Rate of Forgetting and the student's Level of Learning proficiency in interaction with the maximum Level of Learning proficiency, which is 1.

[2] If m = 2, the function is the Logistic curve.

One final point; the probability recall function $p(t) = 1 - e^{-\mu e^{-at}}$ represents a 1-Gompertz formulation, which has a rate of negative RF(t). This is handled in Equation 3 by the negative sign in the integral.

Teaching Curves

Instruction curves (Teaching curves) are an innovation, having no prior conceptualization in education literature. The reasonable choices are between a logistic model of the form $p(t) = 1/(1 - ke^{-gt})$ with rate $dp/dt = gp(t)e^{-gt} = gp(t)(1 - p(t))$, and the **inverse exponential goal seeking function of the form $\boldsymbol{p(t) = 1 - e^{-gt}}$ with $\boldsymbol{goal = 1}$ and rate $\boldsymbol{dp/dt = ge^{-gt} = g(1 - p(t))}$.**

The logistic model would simulate a behavioral pattern where instruction begins slowly, accelerates to a maximum rate and then declines to a zero rate. The initial slow rate of instruction does not seem reasonable considering the Letter Recognition task is the first cognitive task in the reading sequence and teacher's full instruction time can be devoted to it.

The inverse exponential goal seeking model seems a more reasonable behavioral model for instruction, since the rate is at its maximum at t = 0, and decreases to zero as the goal is reached. This simulates behavior where instruction starts at the beginning of the school year at maximum rate for instruction in new letters and decreases as old letters are reviewed and fewer letters of the alphabet remain to be taught. This initial maximum rate of instruction is modeled by the Intensity of Instruction parameter.

Therefore, the Instruction curve chosen is the inverse exponential goal seeking function. The rate for this function is as previously stated. We have this function in Equation 7 with $g = Intensity\ of\ Instruction_{[Race,SES]}$, and P(t)=RPROB1I(t). Equation 7 defines a negative feedback loop between the teacher's Rate of Instruction(RI(t)), and the teacher's Level of Instruction (RPROB1I(t)) in interaction with the teacher's expected proficiency level (Proficiency Goal). We replace 1 by $Proficiency\ Goal_{[Race,SES]}$ because our model formulates the upper limit on teacher's

instruction equal to the teachers expected average proficiency for their students, as moderated by the student's race and SES.

PARAMETER ESTIMATION

Model Initial Conditions

Estimating Initial Level of Learning for the Sample

In order to estimate the values of $Intensity\ of\ Instruction_{[Race,SES]}$, $Proficiency\ Goal_{[Race,SES]}$, and $Initial\ Level\ of\ Learning_{[SES]}$, for all combinations of Race = [African American, White] by SES = [SES1-SES5], it is necessary to fix the value of MFRL and MFRF which are independent of these subscripts. Since these are assumed to be the same on average for all kindergarteners, we 'bootstrap' these values by estimating them from the model over all data ungrouped by race and SES (i.e. [Race, SES] = [RaceAll, SESAll]). But, since it is not possible to estimate Equation 3 for [RaceAll, SESAll], while also estimating its initial condition ($Initial\ Level\ of\ Learning_{[SES]}$ for SES = [SESAll]), we must estimate this value before proceeding.

We estimate $Initial\ Level\ of\ Learning_{[SESAll]}$ from the non-linear Inverse Gaussian fit of Equation 2 over all data. The fit is shown in Figure 13. The estimated parameters are in Table 16. $Initial\ Level\ of\ Learning_{[SESAll]}$, which is $RPROB1L(t)_{[RaceAll,SESAll]}$, at t = 0 is calculated using these estimated parameters with Equation 2, and equals 0.777751.

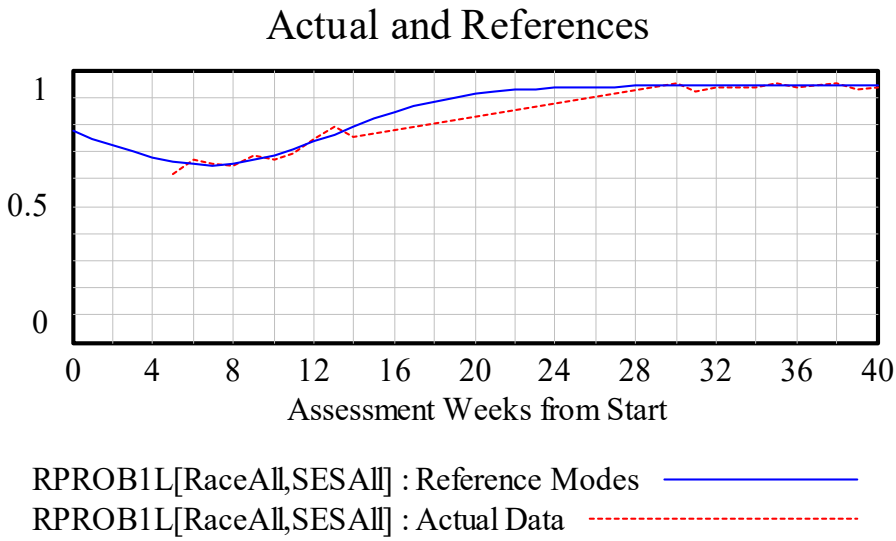


Figure 13. Actual Data and Reference Mode for All Data

Table 16. Inverse Gaussian Fit of RPROB1L for RaceAll, SESAll

Parameter	Estimate	t-score	Prob>t
W	6.2798	11.1052	0.0000
Ymin	0.6505	162.6855	0.0000
Ymax	0.9408	362.5150	0.0000
Xmin	6.7445	15.8105	0.0000

Estimating MFRL and MFRF for the Sample

Using this calculated $Initial\ Level\ of\ Learning_{[SESAll]}$ the model then estimates the values for MFRL and MFRF for [RaceAll, SESAll]. The fit is shown in Figure 14. Parameter estimates are shown inTable 17. Model Parameter Estimates for RaceAll, SESAll Table 17. These values are used throughout subsequent simulations.

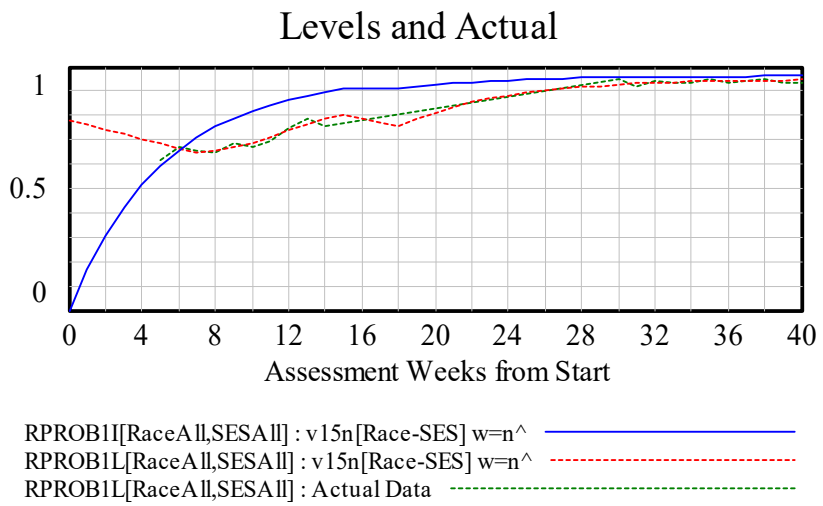


Figure 14. Instruction and Learning Curves for All Data

Table 17. Model Parameter Estimates for RaceAll, SESAll

Parameter	Estimate
MFRF	0.0852689
MFRL	0.337872

Estimation of Model Parameters by Model Reference Optimization

Estimation of SD model parameters can be done by both formal and informal means (Sterman, 2000). Formal estimation methods typically involve some form of regression analysis (linear or non-linear) or non-linear programming techniques. However, Sterman warns that the assumptions underlying regression, particularly OLS regression, typically eliminates OLS as a reliable estimation method for the parameters of SD models, as

> OLS estimates are not accurate in the presence of collinearity (where the variables on the right-hand side are mutually correlated), autocorrelation (where the dependent variable depends on its own past values, that is, where there is feedback), and heteroscedasticity (where the variance of the variables is not constant throughout the sample). (Sterman, 2000, p. 867)

Because of the feedback relationships which typify SD models, collinearity and autocorrelation are characteristic of these models. Additionally, as the processes generating

these variables vary as model structure changes endogenously, model generated data values are typically nonstationary and therefore heteroscedastic.

An additional "caution" is made by Sterman. "[S]tatistical significance of parameters relating variables in an equation is not an indicator of the correctness of the relationship. Statistical significance indicates how well an equation fits the observed data; it does not indicate whether a relationship correctly characterizes causal relationships in the real world." (Sterman, 2000, p. 868). It may therefore be possible to estimate the parameters that will fit an equation without capturing the causal relationships and behavioral processes generating the data. In this case the estimated equation represents the analytical solution to the model, not its structural solution.

The SD software used in this study, Vensim®, offers a robust version of model reference optimization using a non-linear programming search method (Powell's method) with multiple random starting points, to estimate model parameters. Powell's method attempts to estimate model parameters by minimizing the weighted sum of squared errors (min SSE) between the simulated data (X_S) and the actual data (X_A) provided, using a conjugate search algorithm (Powell, 1970). Thus, the Vensim® algorithm attempts to minimize $\sum[(X_S - X_A)w]^2$, subject to the constraints in the model equations, and any parameter constraints set by the modeler. The minimization objective can be modified by application of a weight (w) to the min SSE algorithm. Because the number of observations aggregated as the mean for each time period for the student race and SES subgroups were not equal (see Table 5), the min SSE algorithm was weighted by the square root of the number of aggregated observations, which weighs each time periods aggregate mean by the number of aggregated observations.

Parameter Estimates by Race and SES

The optimization routine was run over the entire model after the estimation of the initial conditions for the Maximum Fractional Rate of Learning (MFRL) and the Maximum Fractional Rate of Forgetting (MFRF) by the 'bootstrap' method previously described.

The estimates for Intensity of Instruction and Proficiency Goal by student race and SES, and for Initial Level of Learning by student SES are shown in Table 18.

Table 18. Model Parameter Estimates by Race, SES

Parameter	Estimate
Intensity of Instruction[African American,SES1]	0.1071
Intensity of Instruction[African American,SES2]	0.1536
Intensity of Instruction[African American,SES3]	0.1773
Intensity of Instruction[African American,SES4]	0.2170
Intensity of Instruction[African American,SES5]	0.2157
Intensity of Instruction[White,SES1]	0.1309
Intensity of Instruction[White,SES2]	0.1791
Intensity of Instruction[White,SES3]	0.2005
Intensity of Instruction[White,SES4]	0.2609
Intensity of Instruction[White,SES5]	0.3410
Proficiency Goal[African American,SES1]	0.9259
Proficiency Goal[African American,SES2]	0.9403
Proficiency Goal[African American,SES3]	0.9655
Proficiency Goal[African American,SES4]	0.9638
Proficiency Goal[African American,SES5]	0.9730
Proficiency Goal[White,SES1]	0.9194
Proficiency Goal[White,SES2]	0.9581
Proficiency Goal[White,SES3]	0.9729
Proficiency Goal[White,SES4]	0.9827
Proficiency Goal[White,SES5]	0.9894
Initial Level of Learning[SES1]	0.5095
Initial Level of Learning[SES2]	0.6704
Initial Level of Learning[SES3]	0.7640
Initial Level of Learning[SES4]	0.8275
Initial Level of Learning[SES5]	0.8966

Vensim® reports that a total of 8787 simulations were run, with 3 random parameter restarts. Each of the 3 restarts converged on the same set of parameter estimates, with the same min SSE, thus giving confidence that the min SSE represented a global rather than local minima. The range for the parameter search for all parameters was constrained to the range 0.0-1.0. Vensim® reported that no parameters 'bump-up' against its constraint during estimation.

The set of parameters show the hypothesized systematic pattern. Intensity of Instruction, which is the initial maximum rate of instruction at the beginning of the school year, shows an expected increase from SES1 through SES5, except for the small (thousandths place) 'inversion' between African American SES4 and SES5, which will be discussed later. The values for Whites are all greater by SES than the value for African Americans. The same systematic pattern is estimated for Proficiency Goal, with the exception of the 'inversion' between African American SES3 and SES4. The deviation from the expected pattern is small, being in the thousandths place. The same expected systematic pattern is seen as the estimated Initial Level of Learning.

The result of applying the estimated parameters to the model can be seen in Figure 15.

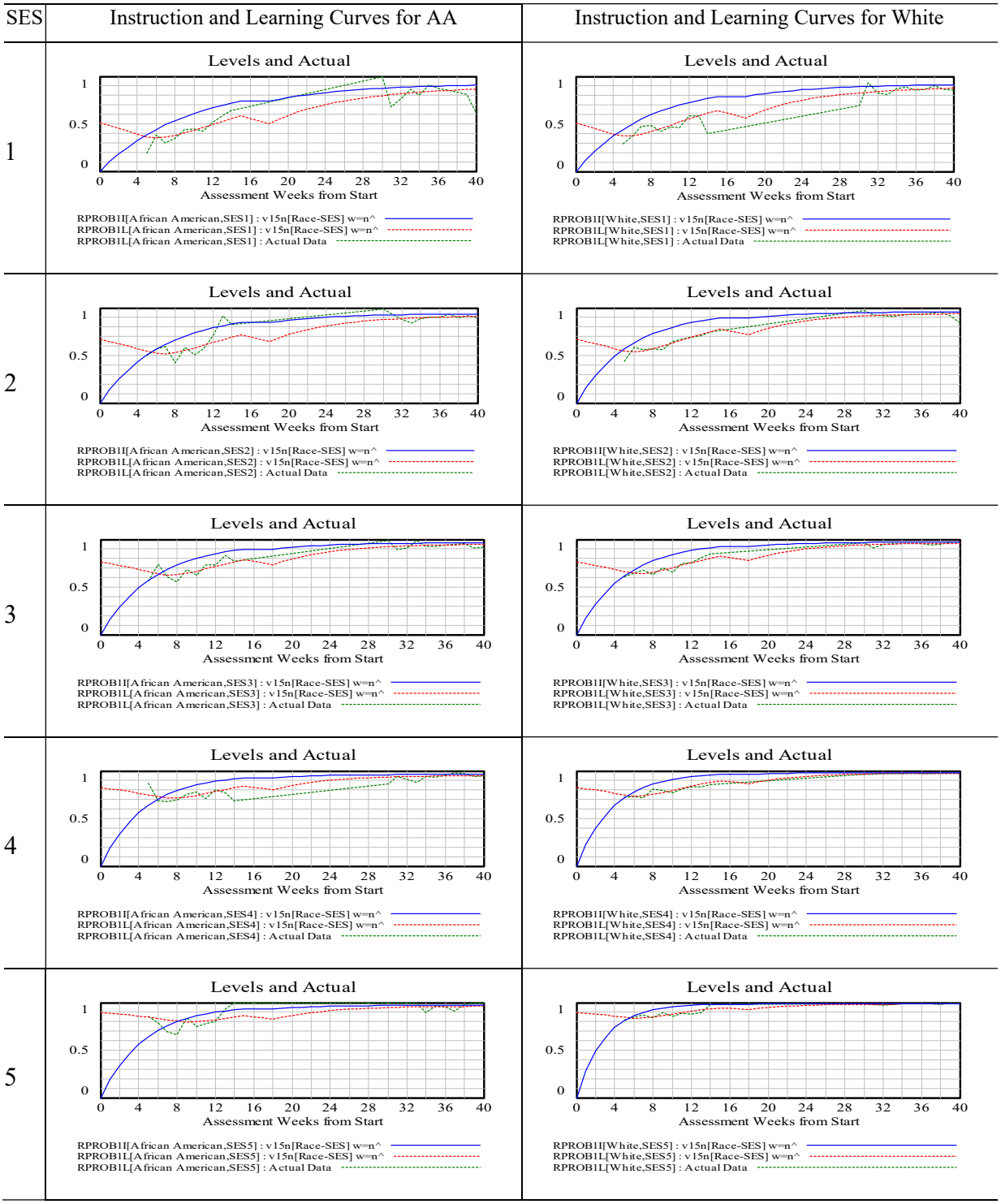


Figure 15. Fit of Instruction and Learning Curves to Actual Data by Race and SES

Visual inspection of indicates a ‘good’ fit for all curves, the exception being the fit for African American, SES5, (enlarged in Figure 16). The fit for African American, SES5 appears to follow a downward trend in the Fall term while the actual data follows an upward trend. A parameter adjustment to generate an ‘improved’ fit and the rationale for this adjustment, as well as a more formal evaluation of ‘goodness of fit’ will be presented below.

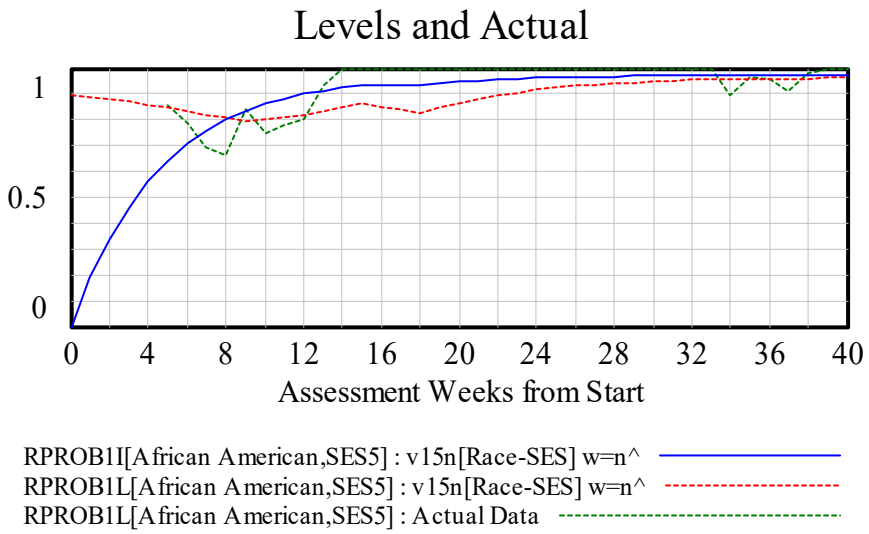


Figure 16. Instruction Curves for African American SES5

The systematic nature of the parameter estimates can be seen by visually inspecting the set of SES by race learning and instruction curves, generated by the parameters applied to the model. This is shown in Figure 17 through Figure 20. Figure 17 shows the generated instruction curves for teachers of African American students by SES. The slopes of the curves, representing “intensity of instruction” increase as we move from SES1 through SES4. We have the ‘inversion’ between SES4 and SES5 (the overlapping dashed lines) but, as indicated, the differences between the curves are negligible and ‘correctible’ with ‘improved’ parameter estimates which generates the top solid line for the ‘improved’ SES5 to be further discussed below. Note also that the slope and Proficiency Goal for SES1 are separated from the SES2-SES5 curves by a noticeable gap.

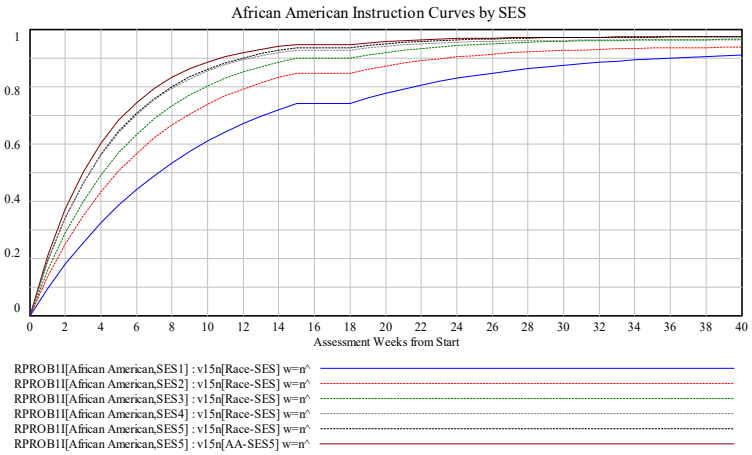


Figure 17.Instruction Curves for African American by SES

The set of generated instruction curves for teachers of White students by SES is as shown in Figure 18. Just as for the African American instruction curves, the slopes increase as we move from SES1 through SES5. There are no "inversions". Again, note the gap between the SES1 curve and the SES2-SES5 curves.

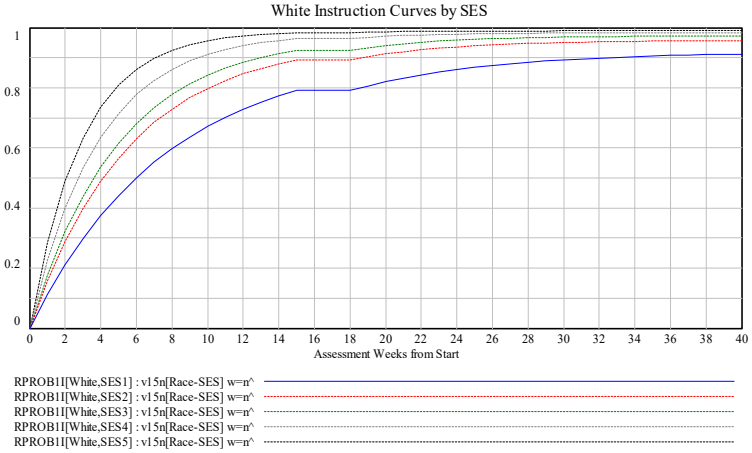


Figure 18. Instruction Curves for White by SES

The generated learning curves for African American students by SES are shown in Figure 19. The curves are obviously layered from SES1 through SES5 (SES5 without 'improvement' is the top dashed line, the 'improved' SES5 is the top solid line) as expected.

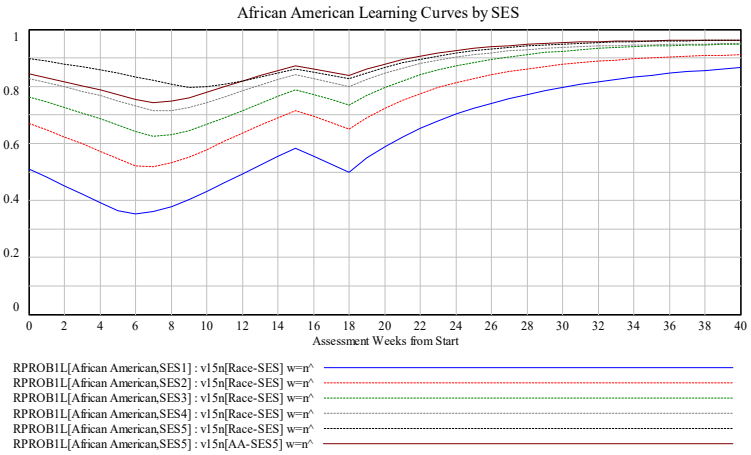


Figure 19. Learning Curves for African American by SES

The learning curves for White students in Figure 20 show the same patterns as the African American curves.

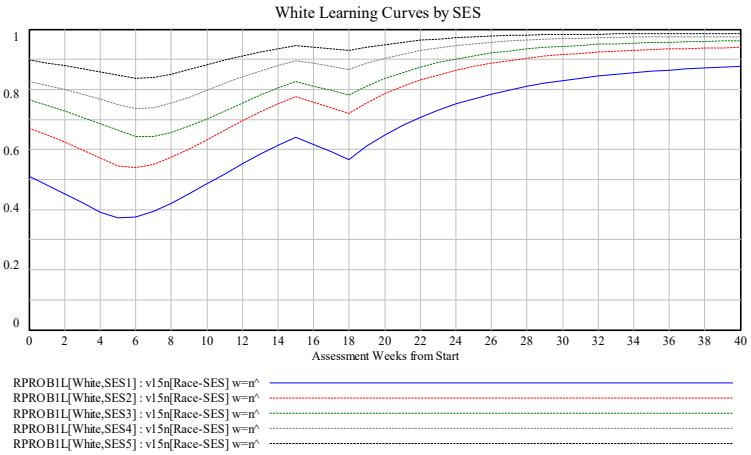


Figure 20. Learning Curves for White by SES

"Improving" the Fit for African American, SES5

Re-estimation of Initial Level of Learning for African American, SES5

The problem in following the trend for the African American, SES5 fit can be seen in Figure 16 for African American SES5. It may appear that the simulated curve has too high a Rate of Forgetting parameter (MFRF) and too low a Rate of Learning parameter (MFRL). However, modifying these parameters to obtain a better fit would violate a

fundamental assumption of the model, i.e. that all children learn and forget, on average, at the same fractional rate. However, by modifying the Initial level of Learning and the Intensity of Instruction for African American, SES5, the fit can be improved. This was done by allowing the model to re-estimate these parameters free of the constraint that the Initial Level of Learning for SES5 is the same for African American and White students. The estimated Proficiency Goal remained the same, as well as all other model parameters. The result of these changes is seen in the "improved' fit in Figure 21, where the simulation better follows the trend of the actual data. The 'improved' parameter estimates are in Table 19.. The Intensity of Instruction 'inversion' between SES4 and SES5 has now been corrected.

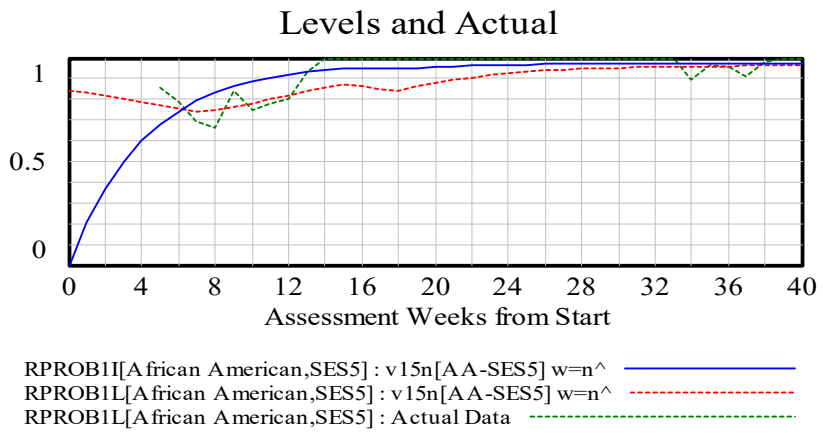


Figure 21. Improved Learning Curve for African American, SES5

Table 19. Improved Parameter Estimates African American, SES5

Improved Parameters	Estimate
Intensity of Instruction [African American, SES5]	0.2407
Initial Level of Learning[SES5]	0.8431

The effect of the 'improved' fit for African American SES5 can be seen in comparison to other learning and instruction curves in Figure 17 and Figure 19. In Figure 17 the 'improved' instruction curve is the top solid line. There now is an 'expected' separation between it and the SES4 instruction curve, systematically so with regards to the

separation for the SES1 through SES4 curves. The 'improved' learning curve in Figure 19 (also the top solid line) has a lowered Initial Level of Learning, closer to SES4's Initial Level, but still distinct from it.

It also now is continuous in the pattern where the marginal difference between each ascending Initial Level of Learning is decreasing. SES, as it makes a difference in Initial Level of Learning, makes less of a difference as SES increases since proficiency probabilities are constrained by an upper value of 1.0. This marginal difference between White SES4 and SES5 is consistent with the hypothesis that there is a greater difference between White SES4 and White SES5 than there is between African American SES4 and African American SES5, to be discussed below.

Justification for Parameter 'Improvements"

We believe these changes can be justified by the following argument. The assumption that children of the same SES, on average, begin school with the same Initial Level of Learning is a reasonable one, supported by research. A higher SES gives one's children better pre-school and home-before-kindergarten 'opportunities to learn'. The SES categories represent quintile categorization of a continuous SES score computed as a composite of parent's occupation, education and income, over the entire sample, which can take the range of -5.0 to 3.0 (National Center for Educational Statistics, 2001b, pp. 7-11). The defined attributes for SES1 through SES4 are the same for all students in these SES categories, African American and White. They have the same upper and lower bounds. But, the SES5 category has no 'true' upper bound. And income levels for White, SES5 are disproportionate at the highest incomes relative to African American, SES5. If the SES attributes for African American SES5 are substantially different from those for White SES5, then the SES differences in Initial Level of Learning for African American and White SES5 students would also be substantially different.

In support of this argument Table 20 presents the minimum, maximum, mean and median values for the continuous SES measure (SESC) for the children in this study, by the quintile SES measure (SESQ5)[3].

Table 20. Continuous SES Statistics by Race and SES Quintile

N	Race	SESQ5	Min SESC	Max SESC	Mean SESC	Med SESC
760	African American	1	-4.75	-0.64	-1.0719	-0.875
727	White	1	-4.47	-0.64	-0.9228	-0.78
546	African American	2	-0.64	-0.31	-0.4767	-0.48
1677	White	2	-0.64	-0.31	-0.4616	-0.46
474	African American	3	-0.31	0.06	-0.1323	-0.13
2012	White	3	-0.31	0.06	-0.1248	-0.12
371	African American	4	0.06	0.61	0.2970	0.27
2331	White	4	0.06	0.62	0.3282	0.33
211	African American	5	0.62	2.27	1.0291	0.93
2814	White	5	0.62	2.67	1.1423	1.05

We can see that the upper bounds for the SES1 through SES4 categories are substantially the same for African American and White students (-.64, -.31, .06, and .61[.62]). However, the upper bound for African American SES5 is 2.27 and for White SES5 it is 2.67. This difference can be expected to differentiate the Initial Level of Learning for African American and White SES5 children. They truly are in different classes. White SES5 have a higher SESC upper bound than African American SES5;

[3] The procedure for creation of the continuous (SESC) and quintile categorical (SESQ5) socioeconomic status (SES) variables is described in the ECLS-K Base Year User's Manual National Center for Educational Statistics. (2001b). *User's manual for the ECLS-K Base Year Public-Use Data Files and Electronic Codebook* (NCES 2001-029 Revised ed.). U.S. Department of Education. . The continuous variable was created as a z-score normalized index computed from father's (or guardian's) education, mother's education, father's occupation, mother's occupation, and household income.

consistent with the hypothesis that there is a greater difference between White SES4 and White SES5 than there is between African American SES4 and African American SES5. Also, note that from the data that African American SES1 have a lower SESC lower bound than White SES1.

DISCUSSION OF RESULTS

Model 'Validation'

A Note on SD Model Validation

Traditional notions of validity are not typically applied to System Dynamics models and are being challenged generally. By traditional notions is meant a position that "validation [is] . . . the process of building confidence in a theory . . . through falsification" (Oliva, 2001, p. 6; Sterman, 2002). Kleindorfer, O'Neill, & Ganeshan (1998) assess the problem of traditional validation as connected with the problem of induction in general and with the idea of strict falsificationism in particular. Any theory, of which a simulation model is one type, seemingly falsified by a counter example, can be saved by the use of "auxiliary hypotheses" (Sterman, 2000, p. 848) or by discounting data (Oliva, 2001, p. 6). That being said, there is no one test that can validate a model. As Sterman points out:

> Models are not validated after they are completed nor by any one test such as their ability to fit historical data. Clients (and modelers) build confidence in the utility of a model gradually, by constantly confronting the model with data and expert opinion-their own and others'. Through this process both model and expert opinions will change and deepen. (Sterman, 2000, p. 81)

Oliva notes the movement of those in the larger Management Science / Operations Research field away from the falsificationist position to what he calls the functionalist position, where the usefulness of a theory serves as its validation criteria. Lane (2001a, 2001b) describes this movement for System Dynamics in particular. Sterman notes that this has been Forrester's viewpoint from the beginning, but until recently the influence of logical positivism ala Friedman (Friedman, 1953) has hindered its acceptance (Sterman,

2002). The incompatibility of the System Dynamics approach to validity, with Friedman's approach becomes apparent when it is understood that System Dynamics models are not instrumental, in the Friedman (or Popperian) sense. The components of a System Dynamics model, levels, rates, positive and negative feedback loops, are intended to "represent the actual nature of physical and social reality" (Forrester, 1994, p. 9). Since these components and how they are related represent the dynamic hypotheses and assumptions of the system being modeled, the closeness to reality is important. This follows naturally since the purpose of System Dynamics modeling is to facilitate policy change (Sterman, 2002) making it essential that "[e]very element of the model should have a real-world counterpart, and every important factor in the real system should be reflected in the model" (Shreckengost, 1985, p. 2). These elements connect observable behavior to hypothesized system structure and decision-making processes (Oliva, 2001). This must be done, cautions Sterman, always understanding that "all models are wrong" to some degree (Sterman, 2002).

As such, System Dynamics assumes neither the extreme objectivist position — "model validation can be divorced from the model builder and its context . . . models are either valid or invalid . . . validation is an algorithmic process which is not open to interpretation or debate," — nor the extreme relativist position -"model and model builder are inseparable . . . all models are equally valid or invalid . . . model validity is a matter of opinion." "[M]ost practitioners have instinctively adopted a middle ground in this debate." (Kleindorfer et al., 1998, pp. 1097-1098), an understanding that facts are important but, it is decision maker's interpretation of the facts that drive decisions not the facts themselves. All facts, not only those useful for formal modeling, but also those in the theories of decision makers are situated in context, and therefore "[v]alidation is the act of increasing to an acceptable level the confidence that an inference about a simulated process is correct for the actual process." (van Horn, 1971, p. 247).

Theil MSE Decomposition Comparison to Actual Data

There are a number of statistical methods for checking the 'goodness of fit' of modeling output to the actual data. Because the distributions of model variables are not known, parametric 'goodness of fit' tests are not suitable. Instead we use a number of non-parametric measures; most notable Theil's Mean Square Error decomposition statistics (Theil, 1966; Theil & Scholes, 1967), as well as Mean Absolute Error (MAE), MAE divided by the mean, and the R^2 statistic.

The Theil statistics were computed using a version of a Vensim module created by Sterman (Oliva, 1996; Sterman, 1984, 2000), modified to handle race and SES subscripting. The Theil statistics decompose mean square error (MSE) into three measures in order to detect systematic differences between actual and simulated data due to unequal means (Um), unequal variance (Us), and unequal covariance (Uc). The decomposition is such that Um + Us + Uc = 1. The statistics are calculated thus: $Um = (\overline{X_s}^2 - \overline{X_A}^2)/MSE$, $Us = ({s_S}^2 - {s_A}^2)/MSE$, and $Uc = 2(1-r)/MSE$. Where $\overline{X_s}$ and $\overline{X_A}$ are the mean of the simulated and actual data, s_S and s_A are the respective standard deviations, and $r = \frac{1}{n}\sum[(X_A - \overline{X_A})(X_S - \overline{X_S})]\,/s_A s_S$, is the correlation coefficient.

According to Sterman, "a large bias, indicated by both a large MSE *and* a large Um" (2000, p. 875, italics in original), indicates a potentially serious systematic problem, "usually due to errors in parameter estimates." While a large Us may indicate a systematic error, such as different trends, if the differences in variation are caused by different amplitudes of the cycle, and if the purpose is to model the trend rather than cyclical behavior, cyclical variation can be ignored. Finally, a large Uc can be problematic if it indicates a phase shift between the simulated and actual data, and the model is intended to capture cyclic behavior, otherwise "if the majority of the error is concentrated in unequal covariation, Uc, the model captures the mean and trends in the data well, differing from the data only point by point" . . . Most likely, a large Uc indicates the presence of noise or cyclical modes in the data series not captured by the model. (Sterman, 2000, p. 877).

Theil Statistics for African American Simulation Runs

An examination of the Um, Us, and Uc statistics for African American students in Table 21 indicates not only a small MSE (less than 1 percentage point squared), but also that for the SES1 through SES4 simulation runs, Uc accounts for at least 90% of the error.

Table 21. Summary Statistics for African American Student Simulations

African American	SES1	SES2	SES3	SES4	SES5	SES5I
MAE	0.0692	0.0558	0.0390	0.0382	0.0536	0.0489
MAE/Mean	0.1096	0.0726	0.0477	0.0453	0.0599	0.0546
MSE	0.0089	0.0064	0.0024	0.0027	0.0046	0.0035
r	0.9104	0.8999	0.9378	0.8808	0.7975	0.8568
R2	0.8287	0.8099	0.8794	0.7759	0.6360	0.7343
Um	0.0175	0.0651	0.0010	0.0479	0.0024	0.0992
Us	0.0386	0.0384	0.0243	0.0334	0.3404	0.0988
Uc	0.9439	0.8965	0.9747	0.9187	0.6572	0.8020

For African American SES5 in the 'standard run', although the mean is captured well, and Uc accounts for 66% of the error, Us is at 34% indicating a problem in following the trend of the data, which was previously noted.

With the improved parameter estimates generating the improved fit, the Theil statistics improve also so that Uc now accounts for 80% of the error (SES5I in Table 21).

Theil Statistics for White Simulation Runs

An examination of the Um, Us, and Uc statistics for White students in Table 22 indicates again not only a small MSE (less than .5 squared percentage points), but for all runs (SES1-SES5), Uc accounts for at least 93% of the error.

Errors statistics are dominated by Uc from early on in the simulation. It is also the case that Um becomes small at or before the 5th data point (week 9) and remains so, indicating that the running mean is captured well by the model. Us is small indicating that trends are captured well.

Table 22. Summary Statistics for White Student Simulations

White	SES1	SES2	SES3	SES4	SES5
MAE	0.0476	0.0227	0.0196	0.0122	0.0152
MAE/Mean	0.0718	0.0291	0.0234	0.0137	0.0162
MSE	0.0048	0.0014	0.0007	0.0003	0.0004
r	0.9466	0.9746	0.9818	0.9834	0.9477
R2	0.8961	0.9499	0.9639	0.9670	0.8982
Um	0.0379	0.0272	0.0064	0.0279	0.0018
Us	0.0333	0.0325	0.0203	0.0372	0.0388
Uc	0.9288	0.9403	0.9732	0.9349	0.9594

Additional Statistical Comparisons

The Theil calculation module also gives an R square statistic (R^2), although Sterman warns that "R^2, though it is widely reported . . . is actually not very useful . . . [since] [t]wo series with the same absolute error can generate very different values for R^2 depending on their common trend." (Sterman, 2000, p. 874). We report the Theil module R^2 along with the Vensim optimization calculated RSquare in Table 23 for comparison. The Theil module R^2 value is the square of the correlation coefficient (r). Sterman prefers this measure since it does not carry the assumption that the means of the simulated and actual data streams are equal. The reported RSquare and R^2 measures are within approximately .01 of each other, and all above .76, except for African American SES5. The RSquare and R^2 values for African American SES5 again confirm the problems with that run. The 'Improved' RSquare and R^2 values for SES5 'improved' was generated by running the model with the 'Improved' parameters generating the fit in Table 23.

As would be expected, the RSquare and R^2 values for White students are higher than those for African American students, reflecting the greater number of observations, and thus the trend curve 'explains' a larger part of the smaller variation in the original disaggregated data streams (refer to Table 5), particularly for SES2-SES5.

Table 23. RSquare and R2 Estimates for Learning Curves by Race, SES

Group	RSquare	R2
RPROB1L[AfricanAmerican,SES1]	0.8256	0.8288
RPROB1L[AfricanAmerican,SES2]	0.7965	0.8099
RPROB1L[AfricanAmerican,SES3]	0.8792	0.8794
RPROB1L[AfricanAmerican,SES4]	0.7637	0.7759
RPROB1L[AfricanAmerican,SES5]	0.6087	0.6360
'Improved' RPROB1L[AfricanAmerican,SES5]	0.7041	0.7343
RPROB1L[White,SES1]	0.8920	0.8961
RPROB1L[White,SES2]	0.9482	0.9499
RPROB1L[White,SES3]	0.9635	0.9639
RPROB1L[White,SES4]	0.9656	0.9670
RPROB1L[White,SES5]	0.8979	0.8982

As a final comparison, the Mean Absolute Errors (MAE), calculated as $MAE = \frac{1}{n}\sum |X_S - X_A|$, and MAE divided by the Mean are reported. The MAEs for the African American runs all fall below .07 points (on a 1.0 point scale). The MAE/Mean reports the MAE as a percentage of the mean and declines from 11% for African American SES1 to 4.5 percent for SES4. This should be the case as the mean proficiency increases from SES1 to SES4. SES5 breaks this pattern rising to 6 percent of the mean. The MAEs for the White runs all fall below .05 points, with most below approx. .02 points. The MAE/Mean declines from approx. 7% for SES1 to 1.3 percent for SES4. There is a slight increase for SES5 to 1.6 percent.

Summary of 'Goodness of Fit"

Do the Theil statistics indicate a good fit of the model output to actual data? Since most of the error is concentrated in Uc, and as the purpose of the model is to capture the trend in Letter Recognition proficiency, and not to account for student variation from the trend or random variation, the presence of large Uc, or larger Us for the African American runs than for White runs, does not lead to a rejection of the model. Sterman writes" [w]hen Uc is large the majority of the error is unsystematic; a model should not be faulted for failing to match the random component of the data" (2000, p. 877).

In summary, we would say that given the overall high R^2 (and RSquare) values, the low MAE and MAE/Mean values, and particularly since the Theil decomposition statistics are dominated by covariance error, rather than mean or variance error, and since the intention is to model the trend in Letter Recognition proficiency rather than to explain variation, the model replicated actual behavior well for our purposes. That being our intention, we want to explore the policy implications for improving Letter Recognition performance, and by that reading performance for African American kindergarten students.

An Unanticipated Validation

What was not anticipated was the higher rate of learning after the drop to the minimum RPROB1L level during Term Break, with a higher rate associated with a lower drop. This phenomenon is completely explainable by what Murre and Chessa describe as 'savings', defined as the relative amount of time saved on the second learning trial as a result of having had the first trial. So, if it takes only half the time to relearn a list, savings will be 0.5. (submitted, p. 9). That is, it takes less time to relearn something already learned (learned and forgotten) than to learn it for the first time. The rate of relearning would be greatest for those students who have forgotten the most, until the original proficiency level is reached. This can be seen in the acceleration in relearning after the Term Break. Again the rate is greater for students who have forgotten the most, and greater forgetting is associated with lower proficiency students since the inverse Gompertz forgetting curve reaches its maximum rate of decline at approximately .368 (Sterman, 2000, p. 299) with progressively lower decline rates going higher and lower. Thus, the African American SES1 curve would have a higher decline rate at Term Break where RPROB1L is approximately .58, than would the African American SES2 curve which has an RPROB1L of approximately .72, and so on through SES5.

This replication of savings in the model occurs because as the student is forgetting, the gap between the Level of Instruction (RPROB1I) is getting larger. The larger the gap between RPROB1I and RPROB1L the greater the rate of learning.

That the model is able to replicate a known empirical phenomenon in learning theory, without an explicit structure to do so, should increase confidence in the robustness of the learning and forgetting parts of the model.

REFERENCE LIST

Anyon, J. (1980). Social Class and the Hidden Curriculum of Work [Article]. *Journal of Education*, *162*(1), 67. http://search.epnet.com/login.aspx?direct=true&db=aph&an=5702077

Behrens, J. T. (1997). Principles and Procedures of Exploratory Data Analysis. *Psychological Methods*, *2*(2), 131-160.

Ferguson, R. F. (2003). Teachers' Perceptions and Expectations and the Black-white Test Score Gap. *Urban Education*, *38*(4), 460-507.

Forrester, J. W. (1994). System Dynamics, Systems Thinking, and Soft OR. *System Dynamics Review*, *10*(2/3), 245.

Friedman, M. (1953). *Essays in Positive Economics*. University of Chicago Press.

Fryer, R. G., Jr., & Levitt, S. D. (2002). *Understanding The Black-white Test Score Gap in the First Two Years of School - NBER Working Paper Series*.

Jussim, L., & Harber, K. D. (2005). Teacher Expectations and Self-Fulfilling Prophecies: Knowns and Unknowns, Resolved and Unresolved Controversies. *Personality and Social Psychology Review*, *9*(2), 131-155. http://www.leaonline.com/loi/pspr

Kleindorfer, G. B., O'Neill, L., & Ganeshan, R. (1998). Validation in Simulation: Various Positions in the Philosophy of Science. *Management Science*, *44*, 1087.

Lane, D. C. (2001a). Rerum cognoscere causas: Part 1 - How do the Ideas of System Dynamics relate to traditional social theories and the voluntarism/determinism debate? *System Dynamics Review*, *17*(2), 97.

Lane, D. C. (2001b). Rerum cognoscere causas: Part II -- Opportunities generated by the agency/structure debate and suggestions for clarifying the social theoretic position of System Dynamics. *System Dynamics Review*, *17*(4), 293.

Lee, V. E., Burkam, D. T., & Economic Policy Institute. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Economic Policy Institute.

Murre, J. M. J., & Chessa, A. (submitted). *A model of learning and forgetting: II. The learning curve*. Retrieved January 28, 2004 from http://www.neuromod.org/staff/murre/drafts

National Center for Educational Statistics. (2001a). *Early Childhood Longitudinal Study - Kindergarten Class of 1998-1999 - Data Files and Electronic Codebook: ECLS-K Base Year Public Use*. In (Version NCES 2001-029 Revised) U.S. Department of Education.

National Center for Educational Statistics. (2001b). *User's manual for the ECLS-K Base Year Public-Use Data Files and Electronic Codebook* (NCES 2001-029 Revised ed.). U.S. Department of Education.

Oliva, R. (1996). *A Vensim® Module to Calculate Summary Statistics for Historical Fit*. Retrieved July 29, 2003 from http://www.sd3.info/models/Library/Misc/TheilStatistics/

Oliva, R. (2001). Model Calibration as a Testing Strategy for Dynamic Hypotheses. Harvard Business School Working Paper 01-047. *forthcoming European Journal of Operational Research*.

Powell, M. J. D. (1970). A Survey of Numerical Methods for Unconstrained Optimization. *SIAM Review*, *12*(1), 79-97. http://links.jstor.org/sici?sici=0036-1445%28197001%2912%3A1%3C79%3AASONMF%3E2.0.CO%3B2-M

Reardon, S. F. (2003, June 2003). *Sources of Educational Inequality: The Growth of Racial/Ethnic and Socioeconomic Test Score Gaps in Kindergarten and First Grade*. Population Research Institute The Pennsylvania State University. Retrieved August 23, 2003 from http://www.pop.psu.edu/general/pubs/working_papers/psu-pri/wp0305R.pdf

Rist, R. (1970). Student Social Class and Teacher Expectations: The Self-Fulfilling Prophecy in Ghetto Education. *Harvard Educational Review*, *40*(3), 411-451.

Rosenthal, R. (1987). "Pygmalion" Effects: Existence, Magnitude, and Social Importance. *Educational Researcher*, *16*(9), 37-41. http://links.jstor.org/sici?sici=0013-189X%28198712%2916%3A9%3C37%3A%22EEMAS%3E2.0.CO%3B2-7

Rosenthal, R., & Jacobson, L. (1968). *Pygmalion in the classroom; teacher expectation and pupils' intellectual development*. Holt.

Saeed, K. (1999). *Defining a problem or constructing a reference mode*. Retrieved September 15, 2003 from http://www.wpi.edu/Academics/Depts/SSPS/Faculty/Papers/13.pdf

Shreckengost, R. C. (1985). *Dynamic Simulation Models: How Valid are They?* Retrieved June 1, 2003 from ftp://sysdyn.mit.edu/ftp/sdep/Roadmaps/RM5/D-4463.pdf

Sterman, J. D. (1984). Appropriate Summary Statistics for Evaluating the Historical Fit of System Dynamics Models. *Dynamica*, *10*(Winter), 51-66.

Sterman, J. D. (2000). *Business dynamics : systems thinking and modeling for a complex world*. Irwin/McGraw-Hill.

Sterman, J. D. (2002). All models are wrong: reflections on becoming a systems scientist. *System Dynamics Review*, *18*(4), 501.

Taris, T. W. (2000). *A Primer in Longitudinal Data Analysis*. Sage Publications.

Tauheed, L. F. (2013). A Critical Institutionalist Reconciliation of “Contradictory” Institutionalist Institutions: What is an Institution?. *Journal of Economic Issues*, *47*(1), 1-22.

Theil, H. (1966). *Applied economic forecasting*. North-Holland Pub. Co.; Rand McNally.

Theil, H., & Scholes, M. (1967). Forecast Evaluation Based on a Multiplicative Decomposition of Mean Square Errors. *Econometrica: Journal of the Econometric Society*, *35*(1), 70-88. http://links.jstor.org/sici?sici=0012-9682%28196701%2935%3A1%3C70%3AFEBOAM%3E2.0.CO%3B2-I

Tobias, B. (2005). *Curve Fitting Functions*. Retrieved May 20 from http://www.originlab.com/pdfs/curvefittingfunctions.pdf

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Pub. Co.

van Horn, R. L. (1971). Validation of Simulation Results [Article]. *Management Science*, *17*(5), 247.