

The Consequences of Standardizing Achievement Scores of Young Children from Low-Income Neighborhoods

by Stephanie Wehry

Introduction

Standardized measures of preschool and primary-grade children's achievement increasingly are used to evaluate school interventions and progress. Frequently, the children are also from low-income neighborhoods. Some national examples include:

Preschool Curriculum Evaluation Research (PCER) • Reading First • Early Reading First Head Start (FACES)

PCER assessments included the *Peabody Picture Vocabulary Test III (PPVT)* and the *Test of Early Reading Ability-3 (TERA-3)*, which have standardized scores and are recommended for use as TIER 1 assessments of language and literacy development and the assessments.¹ Statistical modeling of data typically controls for the age of the children and often results in a negative estimate of the age coefficient. Among young children from low-income neighborhoods, older children in the same grade and within the modal age range:

- Correctly answer the same number or more items.
- Have lower standardized achievement scores.

This negative effect of age on standardized achievement scores of young children experiencing instruction within the same classrooms is counter-intuitive and is inconsistent with findings from the *Early Childhood Longitudinal Study-Kindergarten (ECLS-K)*, which uses proficiency categories and IRT-metrics rather than standardized scores.²

¹ Early Childhood Education and School Readiness Workshop: *Conceptual Models, Constructs, and Measures*. Available at <http://www.niehl.edu/publications/030502workshop.pdf>

² Cooley, Richard J. (2002) *An Unseen Start: Indicators of Inequality in School Readiness*. Princeton, Educational Testing Service. Available at <http://www.ets.org/inequality> and U.S. Department of Education, National Center for Education Statistics, *Entering Kindergarten: Findings from the Condition of Education 2000*, Nicholas Zill and Jerry West, NCES 2001-015, Washington DC: US Government Printing Office, 2001.

Purpose of the Study

The purpose of this study was to investigate graphically and with structural equation modeling the effects of age on standardized and raw achievement scores of young children from low-income neighborhoods.

Sample

In 2002, the Florida Institute of Education at the University of North Florida was awarded a PCER grant to study the effectiveness of a literacy-focused intervention, the *Early Literacy and Learning Model (ELLM)*. The PCER study used a randomized field trial and cluster sampling. Participating classes were selected from low-income neighborhoods in three regions of a southeastern state, each representing a differing degree of urbanicity. Data were collected over a 3-year period.

- 468 4-, 5-, and 6-year-old preschool children in 48 classes were assessed in fall 2002 and spring 2003.
- Slightly less than 200 of these children were assessed as kindergartners in spring 2004 and as first graders in spring 2005.
- Measures included the PPVT Form A, the TERA-3 Form A, and a measure of the children's ability to recognize the 52 upper- and lowercase letters of the alphabet (ABC).³

³ This study used three scale scores from the TERA-3 but did not use the composite Reading Quotient score because of the multivariate design of the statistical models.

The Graphical Study of the Effects of Age

The graphical study used data from the preschool, kindergarten, and first-grade longitudinal study and investigated spring scores from the PPVT and TERA-3 alphabet and meaning scales. These scales were selected because they are widely used and the children's mean scores represent differing levels of achievement. The table below shows PPVT and TERA-3 summary statistics for the scores of the children at each grade level.

PPVT and TERA-3 Summary Statistics from the Longitudinal PCER Study

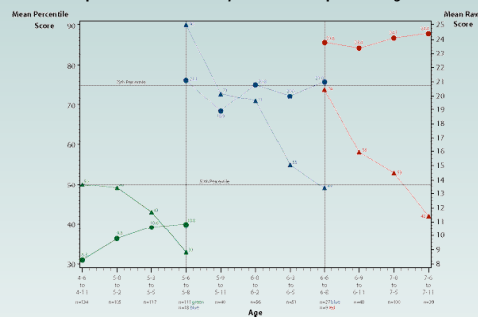
Test	Grade	n	Standardized Scores		Raw Scores	
			Mean	St. Dev.	Mean	St. Dev.
TERA-3 Alphabet Scale	Preschool	184	9.9	3.5	10.4	6.6
	Kindergarten	184	11.3	2.4	20.3	4.4
	First Grade	188	10.2	1.9	24.0	2.2
TERA-3 Meaning Scale	Preschool	184	8.6	1.9	7.3	1.8
	Kindergarten	184	6.6	2.4	9.0	3.0
	First Grade	188	7.5	4.1	16.0	7.1
PPVT*	Preschool	252	90.1	13.4	54.6	16.0
	Kindergarten	187	90.3	12.1	85.0	14.7

Note: *Kindergarten PPVT assessments were not made.

The following figures simultaneously show the percentile rankings of the standardized and raw scores aggregated across age categories of the normative populations for the preschool, kindergarten, and first-grade children. When interpreting the figures:

- Filled triangular markers indicate the mean percentile scores whose scale is on the left vertical axis.
- Filled circular markers indicate the mean raw score whose scale is on the right vertical axis.
- Scores across colors are longitudinal. Green markers/lines represent preschool scores, blue markers/lines represent kindergarten scores, and red markers/lines represent first-grade scores.
- Scores within colors are cross-sectional and represent mean scores of children at the same grade but whose ages at the time of testing fell in different age categories.
- The age category 4-6 to 4-11 means 4 years, 6 months to 4 years, 11 months. Other categories follow the same pattern.

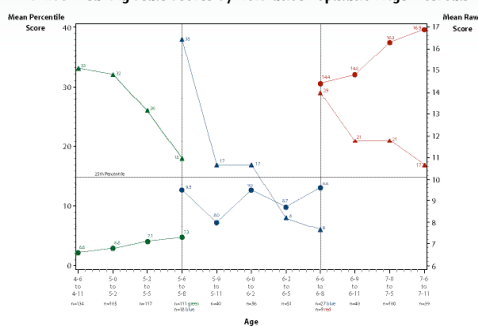
TERA-3 Alphabet Scale Scores by Normative Population Age Intervals



The following observations can be made about the TERA-3 alphabet scale scores:

- Most percentile rankings were above the 50th percentile.
- The preschool mean raw scores increased and the mean percentile rankings decreased across the preschool age categories.
- The kindergarten and first-grade mean raw scores varied around the overall within-grade mean scores and the mean percentile rankings decreased across the within-grade age categories.
- Even though the ages of all children were within the modal age range for their grade, there was an overlap of scores at the grade-level boundary age categories.

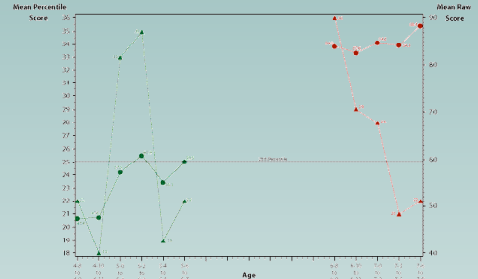
TERA-3 Meaning Scale Scores by Normative Population Age Intervals



The following observations can be made about the TERA-3 meaning scale scores:

- Most percentile rankings were below the 25th percentile.
- The preschool and first-grade mean raw scores increased and the mean percentile rankings decreased across the within-grade age categories.
- The kindergarten mean raw scores varied around the overall mean score and the mean percentile rankings decreased across the kindergarten age categories.

PPVT Scores by Normative Population Age Intervals



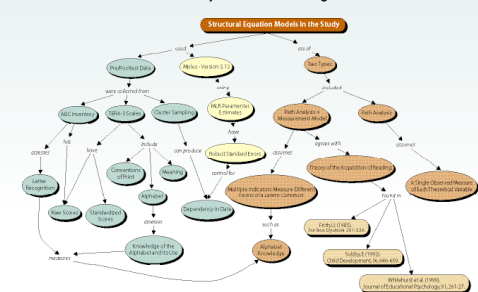
The following observations can be made about the PPVT scores:

- All percentile rankings were below the 36th percentile, and over half were below the 25th percentile.
- There was wide variation in the preschool mean percentile scores, but the mean raw scores showed a tendency to increase across age categories.
- The first-grade mean raw scores showed less variability than the preschool scores and a slight tendency to increase over the first-grade age categories. The mean percentile rankings sharply decreased.

The Structural Equation Modeling Study of the Effects of Age

The figure below depicts the processes used in modeling the data, including information about the assessments used and the models estimated. Blue processes concern data, yellow concern issues specific to model assumptions and statistical software, and brown concern decisions about the types of structural equation models used.

Procedural Map of the Use of Structural Equation Models to Study the Effects of Age



The structural equation modeling used the fall 2002 and spring 2003 data from 468 preschool children nested in 48 classes. The table below shows summary statistics for the raw and standardized scores by the classes' intervention status as ELLM or Control.

TERA-3 Summary Statistics from the 2002-2003 PCER Study of Preschool Children

Test	Metric	Pretest		Posttest	
		ELLM	Control	ELLM	Control
TERA-3 Alphabet	Standardized	7.8	8.0	9.9	9.3
	Raw	4.4	4.8	10.5	9.5
TERA-3 Conventions of Print	Standardized	7.2	7.3	7.7	7.5
	Raw	2.0	2.2	4.5	4.2
TERA-3 Meaning	Standardized	7.8	7.6	8.5	8.0
	Raw	5.5	5.2	7.2	6.7
ABC	Standardized	15.5	13.6	18.0	17.6
	Raw	15.5	33.6	18.0	31.6

Equivalent structural equation models without latent variables (path analyses) and with latent variables (path analyses plus measurement models) were estimated using raw and standardized measures of the children's achievement. The table below provides fit statistics for the models.

Model Fit Statistics

Metric	Estimator	CFI	TLI	RMSEA	SRMR
Standardized	Latent Variables	0.910	0.787	0.129	0.047
	No Latent Variables	1.000	1.002	0.000	0.018
Raw	Latent Variables	0.983	0.962	0.053	0.025
	No Latent Variables	0.995	0.972	0.046	0.039

All models with the exception of the model using standardized scores and latent variables have excellent fit statistics.

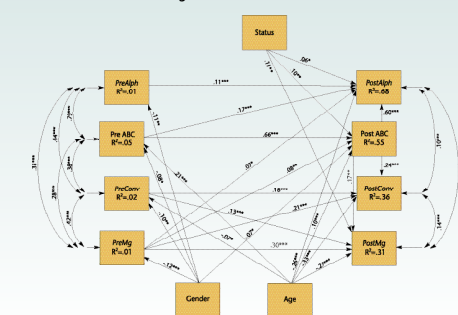
Path diagrams for the path analysis using standardized scores and for the path analysis plus measurement model using raw scores are shown in the figures below, and the following table shows the variable labels used in the path diagrams.

Variables Used in the Study

Variable Label	Variable Description
Gender	Coded 1 for boys and 0 for girls.
Age	Age of the children in months on September 1 of the school year.
Status	Coded 1 for ELLM and 0 for Control.
PreABC (PostABC)	Fall (Spring) ABC score (number of letters recognized).
Alph	TERA-3 Alphabet scale.
Conv	TERA-3 Conventions of Print scale.
Mg	TERA-3 Meaning scale.
PreAlph	Alphabet knowledge fall latent variable measured by PreABC and PreAlph.
PostAlph	Alphabet knowledge spring latent variable measured by PostABC and PostAlph.

Note: Italicized font indicates standardized scores, regular font indicates raw scores. PreAlph, Pre- indicates a fall score, Post- indicates a spring score. Dark orange indicates a latent variable.

Path Diagram - Standardized Variables



- The only direct effects of age on TERA-3 pretest and posttest scores were negative.
- The direct effects of age on the pre/posttest ABC scores were positive.

Information found in the table below allows the comparison of the magnitude of the direct effects of the various independent variables on the spring standardized TERA-3 achievement scores.

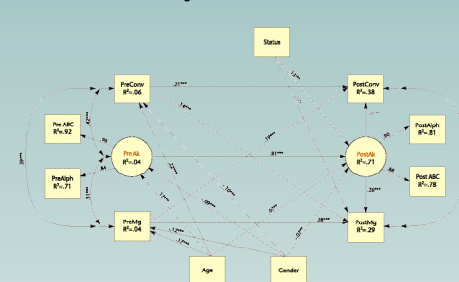
Ranking the Direct Effects of Independent Variables

Dependent Variables	Independent Variables	Standardized Effect	Rank by Size
PostAlph	PreAlph	.110 (.153)	4 (3)
	PostABC	.599 (.559)	1 (1)
	PreABC	.167 (.165)	3 (2)
	Age	-.263 (-.088)	2 (5)
PostConv	PreConv	.184 (.224)	4 (2)
	PreMg	.214 (.205)	3 (3)
	PostABC	.237 (.233)	2 (1)
	Age	-.332 (-.038)	1 (7)
PostMg	PreMg	.296 (.293)	1 (1)
	PreConv	.129 (.146)	4 (3)
	PostABC	.166 (.161)	3 (2)
	Age	-.272 (-.012)	2 (7)

Note: Standardized effect sizes and tests for the raw score variables are in parentheses. * indicates a non-significant effect.

- Age was either the largest or second largest direct effect on all TERA-3 standardized posttest scores.
- Age was not a significant effect on any TERA-3 Posttest raw scores.

Path Diagram - Raw Score Variables



The figure to the left shows the path diagram for the structural equation model of the raw scores.

- There were no direct effects of age on the two TERA-3 posttest scores.
- The direct effects of age on the two TERA-3 pretest scores were positive.
- The direct effects of age on the pre/posttest latent variables were positive.

Conclusions/Discussion

Discussions of the findings of this study concern only the scores and teachers of young children from low-income neighborhoods. Consequently, it is important that future research replicate this study using a random sample of all young children.

Child Development

- If the cognitive development of young children from low-income neighborhoods were the same as that of the children in the national normative sample, age would have a neutral effect on their standardized achievement scores.
- Even though the distribution of the TERA-3 alphabet scale standardized scores approximated the national sample, age remained a significant predictor of the children's emergent literacy achievement.
- The developmental trajectories of the emergent literacy and language abilities of young children from low-income neighborhoods are not the same as the developmental trajectories of the children in the national normative population. Scores suggest the plateaus are broader and the slope is less steep for the children from low-income neighborhoods.

Educational Policy Implications

- When school progress and program efficacy is evaluated using standardized scores, teachers whose children are older – even if the older children are within the modal age range – are at a disadvantage.
- Policy makers should consider what experiences older children from low-income neighborhoods may need in order for their development to keep pace with children of similar age from more affluent neighborhoods.

Test Use

It was expected that the measurement model would fit either score metric because standardization should not change the theoretical construct. The model should fit because:

- ABC assesses children's ability to recognize letters when presented in non-alphabetic order, and all items share the same context – flashcards.
- The first 11 items of the TERA-3 alphabet scale also assess letter recognition, but the context of the items varies from simple to complex contexts that combine letter recognition with the concept of word. The TERA-3 items require higher-order thinking.

The lack of fit of the measurement model provides diagnostic information about the variables and suggests that raw and standardized scores measure different theoretical constructs.

When using standardized measures of achievement of young children from low-income neighborhoods for the purpose of screening children for academic programs:

- Screening for high achievement may over-select younger children.
- Screening for low achievement may over-select older children.

Suggestions

Because highly recommended measures of achievement/ability are widely used to assess young children from low-income neighborhoods, researchers should consider:

- Developing preschool grade-equivalent metrics that are independent of the children's ages.
- Developing preschool and grade-level proficiency categories derived from the test items.
- Developing IRT-based scores.

FLORIDA INSTITUTE OF EDUCATION
Stephanie Wehry, Ph.D.
 Assistant Director for Research
 Florida Institute of Education at the University of North Florida
 For more information call: (904) 620-1197
 email: swehry@unf.edu

The Preschool Curriculum Evaluation Research (PCER) program funded by the Institute of Education Sciences (IES), U.S. Department of Education includes a national evaluation study conducted by RTI International and Mathematica Policy Research (MPR), and complementary research studies conducted by each grantee. The findings reported here are based on the research activities carried out by the Florida Institute of Education at the University of North Florida under the PCER program. These findings may differ from the results reported for the PCER national evaluation study. The content of this presentation does not necessarily reflect the views or policies of the PCER Consortium including IES, RTI, and MPR, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Department of Education.