ESTABLISHING TIME-CONTINUOUS NORMATIVE SCORES FOR TEACHING STRATEGIES *GOLD*® USING A MULTILEVEL GROWTH CURVE MODELING APPROACH

by
Hannah E. Luce

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
In partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Educational Research, Measurement, and Evaluation

Charlotte

2023

Approved by:

____
Dr. Richard Lambert

____
Dr. Kyle Cox

____
Dr. Stella Kim

____
Dr. Drew Polly

ABSTRACT

HANNAH E. LUCE. Establishing Time-Continuous Normative Scores for Teaching Strategies *GOLD*® Using a Multilevel Growth Curve Modeling Approach.
(Under the direction of DR. RICHARD G. LAMBERT)

Young children are assessed to meet federal mandates and inform policy decisions, provide teachers with useful information to make instructional decisions and set reasonable learning goals, and facilitate communication with families. While young children are frequently assessed using whole-child assessments which often yield criterion-referenced score interpretations, norm-referenced score interpretations can help teachers understand relative performance and set reasonable goals for growth. Although researchers have provided validity evidence for both criterion- and norm-referenced score interpretations for one widely used early childhood assessment, *GOLD*®, current national normative scores lack precision for several reasons, including the use of two-time-point and cross-sectional data. To improve estimates, a nationally representative sample of assessment records from 18,000 children ages birth through kindergarten was fitted to a series of hierarchical linear models (HLMs) to establish normative estimates conditioned on months of age and instruction. Secondary study purposes included making inferences about the nature of growth from birth through kindergarten, providing evidence of the most effective time metric for modeling developmental growth, and examining the relationship between child-level characteristics and normative scores. Results indicated that a) HLMs provide reasonably valid normative ability and growth estimates, b) developmental growth, as measured by *GOLD*®, generally slows from ages birth through three years and accelerates from age three through age six, c) the most effective time metric for modeling developmental growth depends on the age band and domain of development, and d) child-level characteristics, including, race/ethnicity, gender, and primary language are associated with

significantly different patterns of preliminary performance and/or growth for children who are one- or two-years of age and older.

DEDICATIONS

I dedicate this dissertation to my husband, Zach, our daughter, Isla, and my parents, Tara, Scott, Dave, and Susie. All of you have always believed in me and provided me with the confidence to see challenges as opportunities for growth. I appreciate your love and support more than you will ever know. And above all else, Isla, I am so proud of you, and I hope you always have the confidence to follow your heart and do what makes you happy.

ACKNOWLEDGEMENTS

Throughout my time in the Educational Research, Measurement, and Evaluation (ERME) program I have been fortunate enough to find not one, but many mentors who have contributed to my success. First and foremost, I would like to thank my academic advisor and dissertation chair, Dr. Richard Lambert. Throughout courses, internships, research assistantships, and the dissertation process you have taught me so much. I appreciate that you have always understood how I learn, and consistently presented me with challenging opportunities, set me free to struggle through them, and offered support as needed. Under your guidance, I have learned to fearlessly accept new challenges, knowing I can always learn the necessary skills later.

Next, I would like to thank my dissertation committee members. First, Dr. Cox, thank you for teaching me how to make complex methodological work accessible and relevant to a broader audience. Next, Dr. Kim, thank you for igniting my interest in measurement through engaging and relevant coursework. Finally, Dr. Polly, thank you for enthusiastically agreeing to serve on my dissertation committee and raising important implications for practice during my proposal and final defense.

Finally, I would like to thank my mentors at NASA, Dr. Amy Chen and Dr. Allison Leidner. Throughout my time at NASA, you pushed me to think critically about the NASA Earth and Space Science Fellowship (NESSF) evaluation framework, dig into the literature, and defend major decision points. Your consistent support led me to become a more independent and critical researcher.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS x

LIST OF EQUATIONS xii

LIST OF FIGURES xiii

LIST OF TABLES xiv

DEFINITIONS xv

CHAPTER ONE: INTRODUCTION 1

- Problem and Purpose 3
- Research Questions 6
- Significance 7
  - Increased Utility of Normative and Growth Scores 7
  - Validity Evidence 9
  - Modeling Time 9
  - Nature of Growth 10
  - Subgroup Differences 10
- Summary of Methodology 11
- Delimitations 12
- Limitations 13

CHAPTER TWO: LITERATURE REVIEW 15

- Assessment in Early Childhood 15
  - Widely Used Authentic Assessment Measures 17
  - Limitations and Directions for Future Research 19
  - Developmental Trends in Early Childhood 21
  - The Nature of Early Development 21
  - Differential Growth Trends 24
- Modeling Growth 29
  - Two-Time-Point Data Analysis 30
  - Cross-Sectional Data Analysis 31
  - Analysis of Variance 31
  - Growth Curve Modeling 32
  - Limitations and Directions for Future Research 36

CHAPTER THREE: METHOD 39

Purpose 39
Research Questions 40
GOLD® 41
Assessment Purposes 42
Constructs 42
Dimensionality 43
Assessment Structure 44
Scores 47
Data Collection Procedures 51
Variables 51
Population 54
Eligibility Criteria 55
Sampling Strategy 56
Sample 57
Data Analysis Procedures 58
Measuring and Modeling Time 58
Data Requirements and Assumptions 60
Software and Estimation Procedures 62
Modeling Typical Development 62
The Most Effective Approach to Modeling Time 68
Inferring the Nature of the Growth Process 69
Examining Subgroup Differences in Growth Trajectories 69
Limitations 75
Ethical Considerations 78
CHAPTER FOUR: RESULTS 79
Hierarchical Linear Model-Estimated Normative Scores 79
Instructional Exposure Norms 79
Age Norms 86
The Most Effective Time Metric 91
Birth to One Year 92
One to Two Years 93
Two to Three Years 94

Preschool, Pre-k, and Kindergarten 94
Patterns Across Age Bands 96
Nature of Developmental Growth 97
Age 98
Instructional Exposure 101
Subgroup Differences 104
Girls 104
Spanish-Speaking Children 109
Children who Speak Other Languages 114
Black Non-Hispanic Children 119
Hispanic Children 123
CHAPTER FIVE: CONCLUSIONS 127
Model-Estimated Normative Scores 127
Implications 129
Recommendations 129
The Most Effective Time Metric 130
Implications 132
Recommendations 133
Nature of Developmental Growth 134
Implications 136
Recommendations 137
Subgroup Differences 137
Girls 138
Children who Speak Spanish 138
Children who Speak Other Languages. 139
Black Non-Hispanic Children. 140
Hispanic Children. 140
Implications 141
Recommendations 144
Summary 146
REFERENCES 148
APPENDIX: DEMOGRAPHIC CHARACTERISTICS OF THE FULL SAMPLE 160

LIST OF ABBREVIATIONS

ANOVA	Analysis of Variance
CFA	Confirmatory Factor Analysis
CFI	Comparative Fit Index
COR	Child Observation Record
COR Advantage	Child Observation Record Advantage
ECLS-K	Early Childhood Longitudinal Study – Kindergarten
ELL	English Language Learner
ESSA	Every Student Succeeds Act
FRL	Free- and reduced-priced lunch
GCM	Growth Curve Model
HLM	Hierarchical Linear Model
ICC	Intraclass Correlation
LGC	Latent Growth Curve
LRT	Likelihood Ratio Test
MANOVA	Multivariate Analysis of Variance
MAP	Measures for Academic Progress
MAR	Missing at Random
MCAR	Missing Completely at Random
ML	Maximum Likelihood
MLM	Multilevel Model
NCLB	No Child Left Behind
PCAR	Principal Components Analysis of Rasch Residuals

REML	Restricted Maximum Likelihood
RMSEA	Root Mean Square Error of Approximation
SEM	Structural Equation Modeling
SRMR	Standardized Root Mean Squared Residual
WHE	Widely Held Expectation

LIST OF EQUATIONS

EQUATION 1: Completely Unconditional Level I Model (Instruction) 63

EQUATION 2: Completely Unconditional Level II Model (Instruction) 64

EQUATION 3: Completely Unconditional Level I Model (Age) 64

EQUATION 4: Completely Unconditional Level II Model (Age) 64

EQUATION 5: Unconditional Level I Instructional Exposure Growth Model 66

EQUATION 6: Unconditional Level II Instructional Exposure Growth Model 66

EQUATION 7: Unconditional Level II Instructional Exposure Growth Model 66

EQUATION 8: Unconditional Level I Age in Months Growth Model 67

EQUATION 9: Unconditional Level II Age in Months Growth Model 67

EQUATION 10: Unconditional Level II Age in Months Growth Model 68

EQUATION 11: Conditional Level I Instructional Exposure Growth Model 71

EQUATION 12: Conditional Level II Instructional Exposure Growth Model 71

EQUATION 13: Conditional Level II Instructional Exposure Growth Model 72

EQUATION 14: Conditional Level I Age in Months Growth Model 73

EQUATION 15: Conditional Level II Age in Months Growth Model 74

EQUATION 16: Conditional Level II Age in Months Growth Model 74

LIST OF FIGURES

FIGURE 1: Mathematics Objective 20 and Corresponding Dimensions a-f 45

FIGURE 2: Physical Objectives 4- 7 and Dimensions 7a and 7b 45

FIGURE 3: Progression for Objective 1a: Regulates Own Emotions and Behaviors 46

LIST OF TABLES

TABLE 1: Instructional *GOLD*® Normative Scores 80

TABLE 2: Reliability of Instructional Intercepts 83

TABLE 3: Comparison of Score Distribution and HLM-Based Instructional Norms 85

TABLE 4: Age *GOLD*® Normative Scores 87

TABLE 5: Reliability of Age in Months Intercepts 90

TABLE 6: Unconditional Model Fit Statistics for the First Three Age Bands 93

TABLE 7: Unconditional Model Fit Statistics for Preschool, Pre-K, and Kindergarten 96

TABLE 8: *GOLD*® Scaled Score Points Per Month Age 100

TABLE 9: *GOLD*® Scaled Score Points Per Month Instruction 103

TABLE 10: Female Effect on Instructional Growth Parameters 107

TABLE 11: Female Effect on Age Growth Parameters 108

TABLE 12: Spanish Effect on Instructional Growth Parameters 112

TABLE 13: Spanish Effect on Age Growth Parameters 113

TABLE 14: Other Language Effect on Instructional Growth Parameters 117

TABLE 15: Other Language Effect on Age Growth Parameters 118

TABLE 16: Black Non-Hispanic Effect on Instructional Growth Parameters 121

TABLE 17: Black Non-Hispanic Effect on Age Growth Parameters 122

TABLE 18: Hispanic Effect on Instructional Growth Parameters 125

TABLE 19: Hispanic Effect on Age Growth Parameters 126

DEFINITIONS

Authentic assessment: is used to "help teachers observe the progress children make through a process that emerges naturally. Evidence of development and learning is gathered during everyday instructional activities" (Lambert, 2020, p. 7).

Balanced data: each participant in the study has the same number of measurements (O'Connell et al., 2008).

Criterion-referenced: "when interpretations are criterion-referenced, absolute score interpretations are of primary interest. The meaning of such scores does not depend on rank information. Rather, the test score conveys directly a level of competence in some defined criterion domain" (American Educational Research Association (AERA) et al., 2014, p. 76).

Developmental assessment: A measure used to understand a child's developmental level in comparison to progressions that demonstrate typical or expected developmental behaviors given age or grade (Lambert, 2020).

Enhanced Assessment Grant: A competitive federal grant program designed to support states in improving the quality of assessment programs designed to measure the academic achievement of elementary and secondary school students (Education Department, 2016).

Every Student Succeeds Act (ESSA): Signed into law in 2015, the Act scaled back the federal government's role in education policy and granted more autonomy to states and schools. However, states and schools were still responsible for establishing accountability systems which included proficiency on state tests, as well as several other indicators (Klein, 2016).

Formative assessment: "a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to help students improve their achievement of intended instructional outcomes" (AERA et al., 2014 as cited in Lambert, 2020, p. 7).

Good Start Grow Smart: Introduced by George W. Bush in 2002, Good Start Grow Smart was an initiative designed to strengthen Head Start, improve early education through partnerships with states, and establish avenues for disseminating knowledge of best practices in early childhood education to teachers, caregivers, and parents (Executive Office of the President, 2002).

Growth curve modeling (GCM): "is a statistical method for analyzing change over time using longitudinal data... Growth curve models focus both on similarities among individuals, captured by the mean structure, and on differences among individuals, captured by the covariance structure" (Diakow, 2018).

Latent growth curve (LGC): An approach to growth curve modeling that can be fitted within a structural equation modeling framework and used to estimate linear or curvilinear growth (O'Connell et al., 2008).

Multilevel model (MLM): An approach to growth curve modeling which includes an extension of the linear regression model and accommodates data with natural or artificial nesting structures (O'Connell et al., 2008).

No Child Left Behind (NCLB): instated in 2002, the law reauthorized the former Elementary and Secondary Education Act and "effectively scaled up the federal role in holding schools accountable for student outcomes" (Klein, 2015, no page number).

Norm-referenced: "when scores are norm-referenced, relative score interpretations are of primary interest. A score for an individual or for a definable group is ranked within a distribution of scores or compared with the average performance of test takers in the reference population (e.g., based on age, grade…)" (AERA et al., 2014, p. 76).

Race to the Top Early Learning Challenge: A federally funded grant program designed to improve the quality of early learning, by a) increasing the number of disadvantaged children in high-quality infant, toddler, and preschool programs, b) designing and implementing a system of high-quality early learning experiences and services, and c) ensuring assessment aligns with research-based practices (Office of Early Childhood Development, 2019).

School Readiness Act of 2003: Passed in 2003, the School Readiness Act amended the previous Head Start Act. The revised Act included many amendments that pertained to funding allocations, community outreach and services, staff qualifications and development, evaluation activities, and the implementation of standards-based learning and assessment activities (H.R. 2210 The School Readiness Act, 2003).

Time-continuous normative score: A normative score that is conditional on variables that change with time, such as age, grade-level, or instructional exposure (Thum et al., 2020).

Time-structured: data are collected at equal intervals for all participants (O'Connell et al., 2008).

Time-varying covariate: An independent variable that may influence the dependent variable and may change over time, e.g., age, instructional exposure or changing schools (O'Connell et al., 2008).

Time-invariant covariate: An independent variable that may influence the dependent variable and does not change over time, e.g., gender, race or ethnicity (O'Connell et al., 2008).

Whole-child assessment: "a balanced way to measure and track a child's progress in all developmental domains… Unlike traditional subject-based assessments, whole-child assessments allow teachers to understand the complete picture of a child" (Marrs, 2020, no page number).

CHAPTER ONE: INTRODUCTION

In 2002, President George W. Bush signed the No Child Left Behind (NCLB) Act which reinstated the previous Elementary and Secondary Education Act. The new legislation was designed to facilitate better outcomes for all students by mandating standards-based education reform and increased accountability through rigorous testing programs (Klein, 2015). The new focus on school- and state-accountability increased the need for tests that measured student achievement and demonstrated how children, schools, and states increased achievement over time. While NCLB was later replaced by Every Student Succeeds Act (ESSA), the use of rigorous and standards-based testing programs and the need to demonstrate continuous improvement remains a cornerstone of the American education system.

Although NCLB was drafted to improve educational outcomes for K-12 students through standardized testing programs and increased accountability measures, early childhood education also experienced significant changes. After NCLB was signed into legislation, a cascade of federal initiatives and grants ensued that enticed states to develop content standards and comprehensive assessment plans for early childhood education programs (Klein, 2016). Policy makers believed that enticing states to develop and implement early content standards, teach language, literacy, and mathematics content, and monitor progress toward rigorous learning objectives, would provide children with the skills needed to reach greater levels of proficiency by the time they participated in state-wide achievement testing programs in third grade and beyond (Stipek, 2006).

While federal-level policies have required states to develop, support, and maintain testing programs, individual states, districts, schools, and teachers have also developed and/or adopted additional measures to understand how young children learn and grow and to inform local-level decisions. Teachers and schools can use ability and growth data to inform instructional decisions, identify strengths and needs (Kohli et al., 2015; Ebel, 1962), monitor progress (Shanley, 2016), set reasonable goals for growth (Thum & Kuhfeld, 2020), and communicate with families and other stakeholders (Burts et al., 2016; Lambert, 2020).

Although achievement and growth data can be useful for many purposes, test scores alone are not meaningful. Instead, teachers, policy makers, and other stakeholders must interpret scores within a given context to make inferences about children. Test score interpretations that involve comparing an individual child's score to a group of scores to determine relative performance can be described as norm-referenced score interpretations (Lok et al., 2016). Norm-referenced interpretations allow for inferences regarding relative standing or growth. Conversely, some test scores are best interpreted by referencing predetermined learning objectives, standards or criterion. These tests are referred to as criterion-referenced assessments, which provide information about a child's level of ability compared to a predetermined benchmark (AERA et al., 2014). While these score interpretations differ, "in practice… there is not always a sharp distinction. Both criterion-referenced and norm-referenced scales may be developed and used with the same test scores if appropriate methods are used to validate each type of interpretation" (AERA et al., 2014. p. 96).

Furthermore, researchers have found that both criterion- and norm-referenced score interpretations can work together to support teaching, learning, and a broader system of accountability (Thum & Kuhfeld, 2020). Criterion-referenced score interpretations can provide teachers with specific and targeted information about each child's unique strengths and weaknesses at the item-level and aid in instructional planning (Ebel, 1962; Kohli et al., 2015), while teachers can use norm-referenced assessment data to understand relative performance and growth over time (Lambert, 2020), and set reasonable learning goals given child-level characteristics (Thum & Kuhfeld, 2020). Together, these unique score interpretations help educational stakeholders support teaching and learning and factor into a broader system of accountability.

Within the context of early childhood education, teachers frequently use criterion-referenced assessments to understand children's knowledge, skills, and abilities across domains of learning and development. Criterion-referenced score interpretations allow early childhood educators to understand each child's current developmental level or academic ability in comparison to predetermined standards. However, criterion-referenced scores do not frequently yield data that allow teachers to meaningfully compare past and present performance, or understand reasonable growth given the child's current level of performance (Lambert, 2020; Thum & Kuhfeld, 2020). Furthermore, criterion-referenced assessments that feature many items or objectives can yield an overwhelming amount of information, making it challenging for teachers to draw inferences about overall performance. Teachers may find additional score interpretations, such as domain-level norm-referenced interpretations, useful in understanding overall performance and what levels of performance are achievable, given child characteristics and patterns of growth over time (Thum & Kuhfeld, 2020).

Problem and Purpose

Teaching Strategies *GOLD*® is the most widely used authentic assessment in early childhood education settings. During the 2021-2022 academic year, over 1.3 million children were assessed using *GOLD*®, including children from all 50 states, D.C., and Puerto Rico. Over time researchers have examined the measurement properties of *GOLD*® scores and established a wealth of validity evidence for the use of *GOLD*® as an authentic, formative, developmental, and criterion-referenced assessment with children ages birth through third grade (Lambert, 2020; Lambert, 2017), including children from diverse linguistic and racial/ethnic backgrounds (Lambert, 2022), and children with disabilities (Kim et al., 2015; Lambert, 2022). However, in a recent technical manual, Lambert (2020) also provided preliminary validity evidence in support of norm-referenced score interpretations. National normative ability and growth scores were established using quartiles, or fourths of the distribution of scores for each age band and major domain of learning or development presented in the assessment system. Although teachers can use the current normative ability and growth scores provided in the manual to understand relative performance, current estimates lack precision and therefore utility, for four primary reasons, including:

1) Lambert (2020) used cross-sectional data to establish ability and growth norms. However, researchers have found that growth norms based on cross-sectional data are not sufficient for estimating intra-individual growth because researchers cannot rule out competing explanations for growth over time (e.g., different curriculum, historical events, or attrition (Singer & Willet, 2003; Thum & Kuhfeld, 2020).
2) Current growth norms were established by taking the difference between pairs of data points, e.g., fall and spring scaled scores. Yet, researchers have found that conducting longitudinal data analysis using two-time-point data is unfavorable when three or more measurement occasions are available (Curran et al., 2010; Singer & Willet, 2003). Researchers have suggested that difference scores, obtained by subtracting the pretest score from the posttest score, may be vastly unreliable, as both pre- and post-test scores are subject to measurement error (Lord, 1956). Furthermore, researchers have argued that the process of change cannot be captured by two-time-point data (Singer & Willet, 2003).

3) The third problem stems from varying assessment windows. Each district or program that uses *GOLD*® has a unique window for finalizing fall, winter, and spring assessment scores. For example, while some programs may finalize fall ratings in early October, other programs may finalize ratings in late December. Currently, normative scores provide teachers with information about average performance at the 25th, 50th, and 75th percentile at the average assessment occasion for fall, winter, and spring for each age band. However, teachers working in programs that finalize assessment scores early may draw the false conclusion that their children are underperforming when compared to their age-level peers. Conversely, teachers working in programs that finalize assessment scores late may draw the false conclusion that their children are outperforming their age-level peers.

4) Finally, Lambert (2020) provided norms by *GOLD*® age bands, which include, birth- to one-year-old children, one- to two-year-old children, two- to three-year-old children, preschool (three- to four-year-old children), pre-k (four- to five-year-old children), kindergarten (five- to six-year-old children), and first grade (six- to seven-year-old children). However, the youngest and oldest children served in each age band may be nearly a year apart. Given that children typically learn and develop rapidly in the early years (Center on the Developing Child, 2016), normative scores provided for one-year age bands that capture average development yield estimates that lack precision for teachers with children who are younger or older than their peers.

Due to the limitations of current national normative ability and growth estimates, the primary purpose of this study was to establish time-continuous normative ability and growth scores for the typical child at each point in time from birth through third grade. By increasing the precision of national normative scores, teachers can make more meaningful comparisons between their children and other children who have received a similar amount of instruction or who are approximately the same age. Furthermore, teachers can use the ability norm estimates to determine expected growth between any two months of age or instruction within the same academic year (Thum & Kuhfeld, 2020). Secondary study purposes included a) identifying the most effective time metric for modeling growth across domains of learning and development for children ages birth through kindergarten, b) making inferences about the nature of the growth process (e.g., shape of the growth trajectory) from birth through kindergarten across domains of learning and development, and c) exploring whether or not and to what extent child-level characteristics are associated with different growth trajectories.

Research Questions

1. What do model-based estimates suggest the average child will do at each point in time from birth through kindergarten, across each of the six major domains of learning and development presented in the *GOLD*® assessment system?
2. Which time metric is most effective for modeling growth, approximate instructional exposure or age in months, from birth through kindergarten?
3. What do hierarchical linear model-estimated slopes from age-separated cohort data suggest about the shape of the developmental pathway from birth through kindergarten across domains of learning and development?
4. How do growth trends differ between subgroups of children (e.g., by race/ethnicity, gender, and primary language)?

Significance

Results from the present study seek to make five significant contributions, including, a) equip teachers and administrators with more precise normative ability and growth estimates, b) provide additional validity evidence for the use of *GOLD*® normative scores with children ages birth through kindergarten, c) establish evidence in support of the best practice for modeling change over time in young children's developmental and academic abilities, d) establish preliminary evidence in support of the best model for demonstrating longitudinal growth from birth through kindergarten across domains of learning and development, and e) provide evidence of differences in growth trends by subgroups of children.

Increased Utility of Normative and Growth Scores

First, the results of this study, including updated tables with time-continuous normative scores for each major domain of learning and development and age band, seek to equip teachers with more precise normative ability estimates. In the presence of more precise estimates, teachers can draw more meaningful conclusions about relative performance given a child's age or instructional exposure. New estimates may be particularly useful for teachers who finalize ratings earlier or later than average or have children who are younger or older than their grade level peers. For example, a teacher who finalizes ratings in December could use the instructional norms to determine the expected domain-level scaled scores for pre-k children after four months of instruction. By using the normative scores corresponding to four months of instruction rather than the average fall assessment occasion, she could obtain estimates that were reflective of the amount of instruction her children actually received. Then, when she compared her children's scores to the estimates corresponding to four months of instruction, she could quickly determine which children may need additional support to meet grade-level expectations.

Similarly, in the presence of more precise estimates, conditioned on months of age and instruction, teachers could also draw more meaningful conclusions about expected growth using any two normative estimates within the same age band. For example, a kindergarten teacher who finalizes fall scores in September and winter scores in January could subtract instructional estimates corresponding to one month of instruction from estimates corresponding to five months of instruction to determine how many scale points, on average, children gain from fall to winter for each domain of learning and development. Next, she could examine how many scale points her children gained between fall and winter and determine which children may need additional support to meet expected growth between January and the end of the year in June.

Finally, administrators can also use normative ability and growth scores to understand aggregate performance and growth trends. For example, a district-level administrator whose pre-k teachers finalized assessment scores in September, October, and November could determine whether average performance across domains of development and classrooms looked similar even though assessment scores were finalized at different points in time. Similarly, administrators whose pre-k teachers finalized fall assessment scores in September and October and spring scores in May, could calculate expected growth for children across seven and eight months of instruction respectively to determine whether children, on average, across classrooms demonstrated patterns of expected growth.

While the examples above are not exhaustive, they highlight how the new time-continuous normative ability and growth estimates could be used to draw more meaningful conclusions about relative performance and growth. Compared to previous estimates which only showcased typical performance at the average fall, winter, and spring assessment occasions, the new estimates provide substantially more flexibility, thus providing more meaningful comparisons.

Validity Evidence

According to researchers, “it is commonly observed that the validation process never ends, as there is always additional information that can be gathered to more fully understand a test and the inferences that can be drawn from it” (AERA et al., 2014, p. 21). Within the context of the present study, evidence of the stability and accuracy of parameter estimates, and subsequent national normative estimates is provided to further substantiate the norm-referenced validity argument. While future research should seek to examine the appropriateness, usefulness, and meaningfulness of norm-referenced score interpretations for children ages birth through kindergarten, the current study posits valid and reliable estimates which are prerequisite to supporting relative score interpretations.

Modeling Time

Additionally, results from the present study provide evidence of the best practice for modeling children’s developmental and academic growth from birth through kindergarten. While researchers have frequently modeled older children’s growth as a function of grade level (Mok et al., 2015; Shanley, 2016; Thum & Kuhfeld, 2020), young children’s growth is closely related to age (Harkness et al., 2013), and changes rapidly (Center on the Developing Child, 2016). Given that young children within the same classroom may vary in age by a full year and development is closely associated with age, modeling ability estimates as a function of age in months may yield more precise estimates (Hujar et al., 2021; Singer & Willet, 2003). Furthermore, while instructional exposure is a significant predictor of academic abilities in mathematics and reading for older children (Thum & Kuhfeld, 2020), researchers have not examined the association between instructional exposure and developmental growth for young children.

Nature of Growth

Furthermore, while the present study used HLMs to model typical developmental growth from the beginning to the end of each academic year, many researchers have fitted multi-year longitudinal data to polynomial models to adequately capture the deceleration of growth over time (Mok et al., 2015; Shanley, 2016; Thum & Kuhfeld, 2020). While this type of analysis requires assessment records from the same cohort of children year after year (Mok et al., 2015; Shanley, 2016; Thum & Kuhfeld, 2020) to rule out competing explanations of growth over time (Singer & Willet, 2003), inferences about the nature of the growth process are made. Results from the present study not only provide evidence of the nature of growth, as measured by *GOLD*® across domains of learning and development from birth through kindergarten, but also may aid in model selection for future research studies using multi-year longitudinal *GOLD*® data.

Subgroup Differences

Finally, *GOLD*® was developed and has been validated for children ages birth through third grade (Lambert, 2020; Lambert, 2017), including children from diverse racial/ethnic and linguistic backgrounds (Lambert, 2022) and children with disabilities (Kim et al., 2013; Lambert, 2020). Researchers have examined and provided evidence of negligible levels of differential item functioning between subgroups of children (Lambert, 2022). Additionally, researchers have provided evidence of measurement invariance between subgroups of children, including, boys and girls, children who are native and non-native English speakers, and White children and children of color (Lambert, 2022). Yet, researchers have not examined how growth trajectories differ by subgroup. Results from the present study contribute evidence of typical growth trajectories by notable subgroups of children, including, boys and girls, children whose primary language is English, Spanish, or *other*, and children who belong to different racial and ethnic groups. Given evidence of invariant measurement between boys and girls, White children and children of color, and native and non-native English speakers, latent means can be compared meaningfully to understand true differences in preliminary development and growth rates over time (Pirralha, 2020).

Summary of Methodology

A series of hierarchical linear models (HLMs) were used to estimate average or typical performance at each point in time from birth through kindergarten across the six major domains of learning and development presented in the *GOLD*® assessment system. Two-level models were used to nest assessment occasions (level one) within individual children (level two). Across models, the outcome variable, Social-Emotional, Physical, Language, Cognitive, Literacy, and Mathematics *GOLD*® domain-level scaled scores, were modeled as a function of either age in months or months of approximate instructional exposure. HLMs were specified for each age band, domain of learning or development, and month of age or instruction, resulting in 72 models (e.g., Kindergarten Unconditional Mathematics Age in Months Growth Model or Kindergarten Unconditional Mathematics Instructional Exposure Growth Model). Next, HLMs were used to understand average developmental level prior to instruction (intercept) and average rate of growth (slope) for every month of instruction and average developmental level for the youngest child in each age band (intercept) and average rate of growth (slope) for every additional month of age. *Pseudo-*r^2 values and Akaike Information Criterion (AIC) compared across non-nested models to determine which time metric accounted for more variation in the outcome, domain-level scaled score, for each age band and domain of development. Next, linear growth rates from age-separated cohorts were examined sequentially to make inferences about the nature of developmental growth, as measured by *GOLD*®. Finally, child-level characteristics, including, gender, race/ethnicity, and primary language, were dummy coded and modeled at level two to examine differences in average primary performance and growth rate over time by subgroup membership.

Delimitations

The primary purpose for the current study was to model typical development at each point in time from birth through kindergarten. Although *GOLD*® is a rater-mediated assessment and researchers have found children's scores are subject to significant rater effects (Hujar et al., 2021), rater-effects were not examined or modeled within the current study. To understand typical growth while minimizing rater-induced nesting effects, two strategies were used. First, only children who shared a rater with no more than 24 other children were eligible to be sampled. Next, frequencies were reviewed to examine how many children in the final sample shared the same rater. Results suggested most teachers in the sample assigned a rating to one child (*n*=9,067, 50.37%), two children (*n*=3,934, 21.86%), or three children (*n*=1,484, 8.24%). While not in the scope of the current study, which focused on average ability and growth estimates regardless of rater behavior, future research should seek to use three level HLMs to nest assessment occasions (level one) within individual children (level two) and children within raters (level three) to understand how rater effects contribute to children's preliminary performance and growth trajectories over time. This type of study would be aided by additional information about raters (e.g., years of teaching experience, level of education, and interrater reliability certification status).

Limitations

The current study used secondary data to a) develop time-continuous normative scores, b) provide evidence of the most effective time metric for modeling young children's developmental and academic growth, c) establish preliminary evidence in support of the nature of the growth process from birth through kindergarten across domains of learning and development, and d) provide evidence of differences in growth trajectories by subgroups of children. However, any time researchers engage in secondary data analysis, where data were collected for other purposes, limitations can arise. Within the context of the current study, there are five data-imposed limitations and one contextual limitation. Data-imposed limitations are summarized below and discussed in greater detail in Chapter Three. The contextual limitation is discussed fully below.

Data-imposed limitations include, a) ability and growth estimates could not be established for first-, second-, and third-grade children, as very few children at these grade levels were assessed using *GOLD*® during the 2021-2022 academic year, b) longitudinal data including no more than three assessments per child placed limitations on the structure of growth models and reliability of model-estimated growth parameters, c) children belong to age-separated cohorts, placing limitations on inferences about the nature of the growth process from birth through kindergarten, d) *GOLD*® data tend to include a greater proportion of low-income children than the general population, therefore limiting the generalizability of normative estimates to the broader population of children ages birth through kindergarten, and e) assessment data for Asian children in birth- to one-year-old and kindergarten classrooms was limited, therefore the sample for the present study slightly underrepresents these populations of children.

The final limitation is contextual and includes the impact of the COVID-19 pandemic. Using 2021-2022 assessment data to establish normative scores may yield atypical results. As a direct result of the pandemic, many children did not have typical early childhood care experiences. Many children were at home with family members, and some preschool and pre-k programs moved to hybrid or online-synchronous instruction. Future research should seek to replicate the current study, using data that are more representative of children's learning and development who have engaged in typical early childhood educational experiences.

CHAPTER TWO: LITERATURE REVIEW

This chapter provides an overview of relevant literature on assessment practices in early childhood, trends in child development and academic abilities, and methods for modeling children's growth over time. A thorough review of extant literature on the topics outlined above provides context for this study which sought to model typical development at each point in time from birth through kindergarten across the six major domains of learning and development presented in *GOLD*®. This research aims to provide teachers with an additional resource to a) make decisions about children, b) set reasonable goals for growth, and c) communicate with families and other stakeholders. Results also aim to inform scholarly methodological literature which is rather void of studies that examine young children's academic and developmental gains using a multilevel modeling approach.

Assessment in Early Childhood

After NCLB was signed into legislation, a series of federal initiatives and grants were created to entice state-level officials to develop and enhance early childhood standards and assessment practices. Federal initiatives, including Good Start Grow Smart and The School Readiness Act called upon states to develop early learning standards in the areas of Mathematics, Language, and Literacy (H.R. 2210 The School Readiness Act, 2003; Stipek, 2006). Additionally, federal grant programs, such as the Race to the Top Early Learning Challenge and the Enhanced Assessment Grant provided states with additional funds to develop comprehensive early childhood assessment systems, including Kindergarten Entrance Assessments (Hanover Research, 2013). Policy makers believed that teachers and schools could boost older student achievement by establishing content standards and systematically measuring academic and developmental progress in early childhood education settings (Stipek, 2006).

In addition to complying with the terms and conditions of federal initiatives and policies, valid and reliable assessment data can also be a powerful tool for individual schools, grade-levels, and teachers. At a local level, assessment data can be used to inform teaching and learning (Hartwig, 2016), set reasonable learning goals (Thum & Kuhfeld, 2020), and communicate with stakeholders (Burts et al., 2016). By understanding each child's unique strengths and needs, teachers can help all children reach their fullest potential.

While there are many reasons that children are assessed in early childhood classrooms, researchers note several significant challenges with assessing young children. First, young children tend to demonstrate their knowledge, skills, and abilities more episodically than older students (Goldstein & Flake, 2015; Wakabayashi et al., 2019). Second, most young children cannot demonstrate their abilities directly using traditional testing formats (Bagnato et al., 2014; Bredekamp & Copple, 2008; National Association for the Education of Young Children, 2020). And third, young children should be assessed across all relevant domains of learning and development, including those that are not typically assessed with older children (Goldstein & Flake, 2015; National Research Council, 2008). Given both the complexity and necessity of understanding how young children grow, develop, and learn, educational researchers have explored numerous approaches to assessing young children and agree that authentic assessment is the best practice (Bagnato et al., 2014; Bagnato & Ho, 2006; Bredekamp et al., 2008; National Association for the Education of Young Children, 2020).

Authentic assessment has been deemed to be the best practice because it allows early childhood educators to observe and systematically document children's knowledge, skills, and behaviors in natural contexts and over multiple occasions to understand what children know and can do (Hartwig, 2016). The structure of authentic assessment allows for teachers to overcome some of the challenges discussed previously by a) gathering evidence of learning and development over time and across different contexts and b) providing children with opportunities to demonstrate their skills and abilities in natural contexts (Bagnato et al., 2014).

Although authentic assessments are widely used in early childhood settings and frequently yield data that are best interpreted using criterion-referenced approaches, in the presence of adequate validity evidence, norm-referenced interpretations can be made as well (AERA et al., 2014; Lambert, 2020). While criterion-referenced score interpretations can provide practitioners with information about children's skills and abilities compared to a predetermined set of benchmarks and aid in instructional planning, they do not frequently facilitate interpretations about growth over time. Norm-referenced score interpretations allow stakeholders to make sense of a child's performance and patterns of growth in comparison to the reference population (AERA et al., 2014). The primary benefits of norm-referenced score interpretations include, a) allowing teachers and policy makers to understand what levels of performance are reasonably achievable, given child characteristics, such as preliminary performance (Reardon & Galindo, 2009), instructional exposure, grade, (Thum & Kuhfeld, 2020), disability status (Hujar et al., 2021), gender (Hujar et al., 2021; Voyer & Voyer, 2014) and primary language (Roberts & Bryant, 2011) and b) providing an accurate depiction of a population of scores (Angoff, 1984). While norm-referenced score interpretations alone are not sufficient for instructional planning, they provide an additional source of information to support teaching and learning and facilitate data-driven decision making at the local-, state- and national-level.

Widely Used Authentic Assessment Measures

To meet federal legislative mandates and provide teachers and schools with information about what children know and can do, test publishers have created numerous authentic formative assessments designed for use with young children. Two of the most widely used authentic assessments include *GOLD*® and the Child Observation Record (COR) or more recently, the Child Observation Record Advantage (COR Advantage). Both assessment systems are observational measures designed to assess the whole child. The first measure, *GOLD*®, was designed and has been externally validated for use with children ages birth through third grade (Lambert, 2020; Lambert, 2017). *GOLD*® includes over 70 items across 10 domains of learning and development: including, social-emotional, physical, language, cognitive, literacy, mathematics, science and technology, social studies, the arts, and English-language acquisition (Burts et al., 2016). The second measure, the COR Advantage, was designed for use with children ages birth through kindergarten and features 34 items across 11 domains of learning and development: including, approaches to learning, social and emotional development, physical development and health, language, literacy, communication, mathematics, creative arts, science and technology, and social studies (Wakabayashi, 2019).

According to The American Educational Research Association (2014), "validity refers to the degree to which evidence and theory support the interpretations of test scores for the proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests" (p. 11). To ensure fair assessment and valid score interpretations, researchers and test publishers have examined the validity of *GOLD*® and COR Advantage scores using empirical data. Over the last two decades, several researchers have sought to establish validity evidence for the use of the COR/COR Advantage as an authentic formative assessment with children ages birth through kindergarten. Researchers have used large samples of children's assessment records to examine and provide evidence of construct validity (Akaeze et al., 2022; Fantuzzo et al., 2002; Wakabayashi et al., 2019), content validity, concurrent validity, and interrater reliability (Wakabayashi et al., 2019). Although researchers have established a substantial amount of validity evidence to support the use of the COR Advantage with young children, none of the studies used nationally representative samples of children, therefore limiting the generalizability of results.

In addition to providing validity evidence for the use of the COR/COR Advantage with young children, researchers have also sought to provide validity evidence for the desired interpretations of *GOLD*® scores. Over time, researchers have used nationally representative samples of assessment records to provide validity evidence for the use of *GOLD*® with children ages birth through third grade (Lambert, 2020; Lambert, 2017), including children with disabilities and English Language Learners (ELLs) (Kim et al., 2013; Lambert, 2022). Researchers have examined (Lambert et al., 2015) and reexamined internal structure over time to provide evidence of construct validity (Lambert, 2020). Additionally, preliminary evidence has been established for criterion validity (Lambert, 2020), concurrent validity (Kim et al., 2013), measurement invariance across assessment occasions within one academic year (Lambert et al., 2015), and measurement invariance across subgroups of children, including gender, race/ethnicity, disability status, and primary language (Lambert, 2022). While the majority of validity evidence has been collected to support the use of *GOLD* ® as an authentic, formative, and criterion-referenced assessment, researchers have also provided preliminary evidence for norm-referenced interpretations of *GOLD* ® domain-level scaled scores (Lambert, 2020).

Limitations and Directions for Future Research

Young children are assessed for numerous purposes, including, a) to meet accountability mandates, b) provide teachers with useful information, and c) communicate with families and other stakeholders. While amassing validity evidence and substantiating validity arguments should be prerequisite to interpreting scores and making decisions about children, Goldstein et al. (2015) suggest that many assessments designed for use with young children have been subject to less rigorous validation research than widely used measures designed for adolescence and adults. After careful review of studies that sought to provide validity evidence for two of the most widely used authentic assessments in early childhood classrooms across the U.S., it was apparent that significant gaps exist. None of the studies conducted with COR or COR Advantage assessment records used nationally representative samples of children. When validation research is conducted without representative samples of children, it can be challenging, if not impossible, to understand whether scores can be extended to all subgroups of children (AERA et al., 2014). Future research should replicate earlier study methods using nationally representative samples of children to ensure results can be generalized to the broader population of children assessed with the COR Advantage.

Furthermore, while studies conducted to examine the validity of *GOLD*® scores frequently used nationally representative samples of children's assessment records, additional evidence should be provided in support of interrater reliability and concurrent validity. Additionally, "when test scores are interpreted in more than one way… each intended interpretation must be validated" (AERA et al., 2014, p. 11). Although there is a significant amount of validity evidence to support criterion-referenced score interpretations, evidence to support norm-referenced interpretations is limited. Lambert (2020) provided normative ability and growth estimates across the six major domains of learning and development included in the *GOLD*® assessment system for children ages birth through first grade. Yet, estimates lack precision due to the methods used. Future validation studies should seek to establish more precise norms using methods that leverage between-child variability in assessment dates and children’s ages at the time of assessment.

Developmental Trends in Early Childhood

For decades educational stakeholders have been interested in understanding how young children develop and learn over time. Understanding how children grow allows policy makers to craft informed and relevant legislation and for test developers and curriculum writers to create developmentally appropriate standards, content, and assessments. Similarly, when teachers recognize how most children develop and acquire new skills and abilities, they can tailor curricular materials and instruction to meet children’s unique strengths and needs.

The Nature of Early Development

From ages birth through five children’s brains develop more rapidly than at any other point in the lifespan (Harkness et al., 2013; Rebello-Britto et al., 2013). Furthermore, researchers have observed that some level of growth and development is biological (Harkness et al., 2013), while other aspects of development are profoundly influenced by ecological and environmental factors (Chetty et al., 2011; Gialamas et al, 2013). For example, researchers have widely observed that babies, regardless of geographic location or beliefs about caretaker interactions, reach developmental milestones, such as crawling, walking, and talking at similar points in time (Harkness et al., 2013). Yet other researchers have found experiences with primary and secondary caregivers and the environment interact and influence development (Gialamas et al., 2013; Yoshikawa et al., 2013). For example, in one study, Gialamas et al. (2013) found that higher quality relationships between caregivers and two- and three-year-old children were associated with greater cognitive abilities and social-emotional competence at ages six and seven, after controlling for child and parent characteristics. Similarly, Chetty et al. (2011) found that children with an excellent kindergarten teacher, rather than poor- or average-quality kindergarten teacher, were more likely to receive higher ratings on soft skills by their eighth-grade teachers and earn significantly more money over a lifetime. Together, these studies highlight the complexity of early childhood development and the dynamic relationships between biology, caregiver interactions, and the environment where children learn and grow.

To simplify the complexity that surrounds development in early childhood, stakeholders frequently deconstruct early childhood learning and development into discrete domains (e.g., language, cognitive, physical, social-emotional, etc.). Yet, there is growing consensus that young children's learning is highly interconnected and interrelated across domains of learning and development (Lambert, 2020, Pace et al., 2019; Rebello-Britto et al., 2013). For example, as a child develops greater receptive language skills, she may also use a wider variety of words while speaking. Similarly, a child who develops greater spatial reasoning may also develop more complex gross motor skills. While learning and development in one area often positively influences other areas, researchers have also observed how children can experience uneven development across tasks of equal cognitive complexity (Piaget, 1971). For example, a child may use conventional grammar in speech, but not in writing. Although individual children may experience some unevenness in the acquisition of skills, it's likely that data aggregated to the classroom- or school-level will demonstrate steady growth trajectories across individually measured domains or objectives of the same cognitive complexity (Lambert, 2020).

Another trend that stems from the interconnectedness of early learning and development is that social-emotional competence in early childhood is positively associated with later academic achievement (Blewitt et al., 2018; Corcoran et al., 2018; Pace et al., 2019; Rhoades et al., 2011) and early academic skills are positively associated with later academic achievement

(Goldstein et al., 2015; Kuhfeld et al., 2020; Pace et al., 2019). For example, Rhoades et al. (2011) examined the relationship between social-emotional competence at the beginning of preschool and academic achievement in first grade. Results suggested that social-emotional competence in preschool was a significant predictor of academic ability after controlling for the effects of maternal education, household income, and child demographic variables. Likewise, Pace et al. (2019) examined the predictive relationship between language, reading, mathematics, and social-emotional skills at school entry and later elementary grades. Results suggested that ability at school entry within each discrete domain of learning or development was the greatest predictor of ability at successive measurements within the same domain. Additionally, language, reading, and mathematics abilities at school entry were predictive of later achievement within other domains as well. Together these studies highlight the inseparable nature of early childhood development and academic achievement in elementary school.

Although researchers have found children who demonstrate stronger developmental (Rhoades et al., 2011; Pace et al., 2019) and academic (Pace et al., 2019) abilities early on often continue to outperform peers in later grades, researchers have also observed that growth rate, regardless of preliminary status, tends to decelerate over time across domains of learning (Lee, 2010; Mok et al., 2015; Shanley, 2016). For example, in one study, Lee (2010) examined growth trends in mathematics and reading using a nationally representative and cross-sectional sample of Comprehensive Test of Basic Skills (CTBS) assessment records for first through twelfth grade students. Results suggested that children made greater gains in reading and mathematics early on and fewer gains as time passed. Similarly, Shanley (2016) used latent growth models to examine trends in mathematics growth using the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K) data. Piecewise models demonstrated the best fit due to the decelerating growth rate in mathematics from kindergarten through eighth grade. Together, these studies illustrate the nature of children's growth which is often rapid early on and slows as children progress through subsequent grade levels.

Differential Growth Trends

In addition to exploring aggregate developmental and academic growth trends, researchers have also sought to understand how specific demographic characteristics are associated with children's abilities before kindergarten (Hujar et al., 2021; Miller et al., 2021), at school entry (Kuhfeld et al., 2020; Reardon et al., 2016; Roberts & Bryant, 2011) and beyond (Hujar et al., 2021; Mok et al., 2015; Reardon et al., 2016; Reardon et al., 2009; Shanley, 2016; Voyer & Voyer, 2014). While operationalizing, measuring, and isolating child-level variables can be a difficult task (Roberts & Bryant, 2011), identifying influential child-level characteristics allow policy makers to write legislation to promote equity and close opportunity and achievement gaps (Hudson, 2015; Kuhfeld et al., 2020).

One child-level characteristic that has been explored extensively is family income. Researchers have examined the relationships between children's developmental and academic abilities and a variety of variables that approximate family income, such as, socio-economic status (Roberts & Bryant, 2011), FRL status (Hujar et al., 2021), and parental income (Reardon & Portilla., 2016). Research suggests that children from low-income families tend to demonstrate lower abilities early on (Hujar et al., 2021; Reardon & Portilla, 2016) and slower growth rates over time (Hujar et al., 2021; Roberts & Bryant, 2011). For example, Reardon & Portilla (2016) examined ECLS-K math and reading scores for children from low-, middle-, and high-income families and found that on average, scores were higher at each successive level of income. Additionally, while achievement gaps between children from wealthy and poor families narrowed significantly from 1998 to 2010, differences in ability at kindergarten entry continued to persist by family-income level (Reardon & Portilla, 2016). Furthermore, Roberts and Bryant (2011) examined differences in mathematics ability over time given children's socioeconomic status and home. Results suggested low socioeconomic status was the most significant predictor in the model and was strongly associated with slower rates of growth across over time, regardless of home language. Together these studies suggest that poverty is significantly related to lower abilities early on and slower rates of growth over time.

In addition to family-income level, researchers have also examined how ability and growth trends differ by race, ethnicity, English-Language-Learner (ELL) status, and primary language. Yet, understanding the ways in which these child-level characteristics are related to abilities in early childhood and subsequent growth trajectories is complex. Although researchers frequently conduct subgroup analyses to understand how abilities and growth trajectories differ by child-level characteristics, groups often consist of heterogeneous populations (Roberts & Bryant, 2011). For example, researchers may examine mathematics ability at school entry between native-English speakers and ELLs. However, the children assigned to the ELL group may include children at different levels of English-language proficiency and children from different linguistic and cultural backgrounds. Similarly, researchers may be interested in understanding differences in literacy ability between White and Black children, yet ethnic subgroups may exist within racial categories. While examining differences by race, ethnicity, home language, and ELL status can be challenging, researchers have identified several trends that yield important information for researchers, practitioners, and policy makers.

To begin, researchers have found that ELLs and children who speak languages other than English at home tend to demonstrate different abilities in early childhood (Hujar et al., 2021) and different rates of growth over time (Hujar et al., 2021; Roberts & Bryant, 2011). For example, in one study, researchers found that ELLs demonstrated lower abilities in mathematics, literacy, language, social-emotional, cognitive, and physical domains in the fall semester of Pre-K, yet experienced stronger rates of growth than their English-fluent peers between fall and spring measurements (Hujar et al., 2021). Furthermore, Roberts and Bryant (2011) examined the relationship between children who were labeled English-language proficient, but spoke various languages at home, and trends in mathematics and reading achievement. Results suggested that children who spoke Spanish at home were more likely to demonstrate lower abilities at kindergarten entry than children who spoke English or an Asian language at home. Furthermore, measurements in kindergarten for Spanish-speaking children were more predictive of subsequent measurements, suggesting it was more difficult for Spanish-speaking children to overcome gaps that were present at school entry (Roberts & Bryant, 2011). Together, these studies further illuminate the complexity of understanding relationships between both English-language proficiency and primary language and abilities in early childhood and subsequent growth rates.

While disentangling race and ethnicity from other influential child-level characteristics can also present challenges (Roberts & Bryant, 2011), researchers have identified several notable trends between race and ethnicity and academic and developmental growth. Recent research suggests that achievement gaps have narrowed across academic and most developmental domains at school entry between White children and Black children (Kuhfeld et al., 2020) and White and Hispanic children (Reardon & Portilla, 2016; Reardon & Galindo, 2009), yet gaps continue to persist to some extent (Kuhfeld et al., 2020; Reardon & Portilla, 2016). Furthermore, over thirty years of National Assessment for Educational Progress (NAEP) data suggests gaps between Black and White children don’t dissipate after school entry, but in fact persist through high school (Young et al., 2017). And, while Hispanic children may experience more rapid growth than their White peers early on, growth rate decelerates over time and gaps continue to persist through later grades (Reardon & Galindo, 2009). Together these research findings describe observable patterns of White Non-Hispanic children demonstrating stronger academic and developmental skills and abilities at school entry than Black and Hispanic children, and stronger rates of growth over time.

Finally, researchers have explored differences in preliminary performance and growth over time between boys and girls. Research tends to suggest that girls demonstrate stronger abilities at school entry (Kuhfeld et al., 2020; Voyer & Voyer, 2014) that persist over time (Voyer & Voyer, 2014). For example, researchers examined kindergarteners' Measures for Academic Progress (MAP) Reading and Mathematics Growth assessment records to understand whether boys or girls demonstrated greater abilities at school entry. Results suggested that girls demonstrated stronger abilities in math and reading at school entry across seven years of assessment records (Kuhfeld et al., 2020). Furthermore, to understand if gender gaps persisted throughout K-12 education and beyond, Voyer & Voyer (2014) conducted a meta-analysis using the results from over 300 studies. Results suggested that across grade levels and subjects, regardless of scoring system, girls tended to outperform boys. Yet, this trend isn't upheld across all studies and contexts. In another study conducted by Hujar et al. (2021), boys' and girls' *GOLD*® mathematics, literacy, language, cognitive, physical, and social-emotional assessment records were compared to understand differences in preliminary ability in Pre-K and growth throughout the pre-k year. Results suggested that boys outperformed girls at the beginning of pre-k in the areas of literacy, math, language, and cognitive abilities. Yet, girls demonstrated stronger rates of growth across most domains, including literacy, math, physical, and social emotional. While research tends to suggest girls outperform boys across subjects and time, this finding is not upheld across all studies and contexts.

Limitations and Directions for Future Research

While many studies have found that caregiver interactions and early childhood program quality have a profound influence on development (Chetty et al., 2011, Gialamas et al., 2013; Yoshikawa et al., 2013), other researchers suggest that infants and young toddlers, regardless of geographic location, environment or beliefs about caregiver-child interactions, reach many developmental milestones at similar points in time (Harkness et al., 2013). Future research should seek to examine the relationships between instructional exposure and age and subsequent growth to understand whether environmental and ecological factors or biological age have a greater influence on developmental growth at each point in time throughout early childhood.

Additionally, while researchers have widely observed that growth rate decelerates over time across domains of learning and development (Lee. 2010; Mok et al., 2016; Shanley, 2016; Shin et al., 2013), average gain score estimates provided in the Technical Manual for the Teaching Strategies *GOLD*® manual do not reflect this pattern of growth (Lambert, 2020). Instead, gain scores, which reflect the average difference between fall and spring measurements for each domain and age band, suggest that children in birth- to one-year-old classrooms and kindergarten classrooms make the most significant gains. If gain scores were consistent with the broader literature, we would expect to see diminishing gain scores as children progress from birth through first grade, yet this is not the case. Future research should seek to use growth curve modeling methods to derive more precise growth estimates. Model estimates should be reexamined to determine whether developmental and academic growth, as measured by *GOLD*®, provide further evidence in support of diminishing growth rate over time.

Finally, many researchers have sought to understand how demographic variables, such as family-income, race, ethnicity, home language, and gender relate to academic and developmental performance and growth in early childhood. While research generally suggests that children living in poverty (Hujar et al., 2021; Reardon & Portilla 2016; Roberts & Bryant, 2011), Black children (Kuhfeld et al., 2020; Young et al., 2017), Hispanic children (Reardon & Portilla, 2016; Reardon & Galindo, 2009), children who speak Spanish at home (Roberts & Bryant., 2011), and boys (Kuhfeld et al., 2020; Voyer & Voyer, 2014) tend to demonstrate lower abilities in early childhood and slower rates of growth, child-level demographic variables can be challenging to isolate, operationalize, and measure (Roberts & Bryant, 2011). Furthermore, although researchers often seek to establish homogenous subgroups, such as ELLs or Black children, within-group heterogeneity and intersecting identities uniquely influence outcomes (Roberts & Bryant, 2011). Although the literature can serve as a guidepost for researchers who seek to examine and model the relationships between academic performance, developmental progress, and a variety of child-level characteristics, it also presents a somewhat unclear picture filled with nuances. Researchers must decide how to operationalize child-level characteristics, set criteria for inclusion and exclusion, determine how developmental progress and academic abilities will be measured and modeled, and place cautionary limitations on findings.

Modeling Growth

Although researchers have been interested in modeling change over time for generations, NCLB established a stronger need for educational researchers to model change in student achievement (Thum & Kuhfeld, 2020). While there are many different approaches to modeling change over time, some are better than others, especially within the context of large-scale assessment data, where children are not assessed on the same schedule and do not always share the same number of measurements (O’Connell & McCoach, 2008; Singer & Willet, 2002; Thum & Kuhfeld, 2020).

Before the development and widespread use of growth curve models, researchers frequently analyzed the change between pairs of measurements (Bryk & Weisberg, 1976; Curran et al., 2010; Rogosa & Willet, 1983). To examine differences between pairs of observations and understand intra- and inter-individual change over time researchers frequently used three methods, including, two-time-point data analysis, cross-sectional data analysis, and Analysis of Variance (ANOVA) or Multivariate Analysis of Variance (MANOVA) (Curran et al., 2010; O’Connell & McCoach, 2008; Singer & Willet, 2003). While all of these methods are still used today and may be appropriate for smaller-scale studies where researchers wield significant control over data collection procedures, none of these methods are sufficient for understanding intra- and inter-individual growth when data collection schedules and the number of measurements per child vary vastly from program to program and state to state (Thum & Kuhfeld., 2020).

Two-Time-Point Data Analysis

Although pretest-posttest designs are frequently used in educational contexts to quantify the amount that children learned from one point in time to the next, researchers have argued that difference scores, obtained by subtracting the pretest score from the posttest score, are vastly unreliable (Lord, 1956). Lord (1956) argued that within the context of educational and psychological measurement, both pretest and posttest scores are subject to measurement error. Therefore, difference scores are even less reliable than single measurements. Singer & Willet (2003) also criticized the use of two-time-point data analysis procedures, stating that any researcher who uses such methods “narrowly conceptualize[s] change as an increment: the simple difference between scores assessed on two measurement occasions" (Singer & Willet, 2003, p. 10). In other words, Singer & Willet (2003) argued that a simple difference score could not adequately capture the nature of the growth process.

Cross-Sectional Data Analysis

Researchers have also used cross-sectional data to infer how individuals change over time (Singer & Willet, 2003; Lambert, 2020). However, researchers have suggested that using cross-sectional data is insufficient for understanding intra-individual growth because differences in observed outcomes between measurements could be due to true growth or other factors such as attrition (Singer & Willet, 2003; Thum & Kuhfeld, 2020).

Analysis of Variance

Researchers have also used ANOVA and MANOVA to make inferences about growth over time (Curran et al., 2010; O'Connell & McCoach, 2008). However, researchers working from an ANOVA framework face significant constraints in terms of data structure, including, a) data must be balanced and b) data must be time-structured (O'Connell & McCoach, 2008; Singer & Willet, 2003). In other words, the number of measurements must be the same for all individuals and all individuals must be assessed on the same schedule. If individuals are missing datapoints, then the researcher must exclude the case from analysis (Curran et al., 2010). Similarly, if the individual was assessed at a different point in time, the researcher must listwise delete the case or accept the difference in time as a source of random error (O'Connell & McCoach, 2008). Although listwise deletion allows the researcher to proceed with analysis, deleting participants can bias estimates and negatively impact the power to detect effects (Raudenbush & Bryk, 2002).

Growth Curve Modeling

In recent decades, researchers have turned to growth curve modeling to combat the limitations of earlier methods of analysis and allow for greater flexibility, including the accommodation of partially missing data (O'Connell & McCoach, 2008), non-linear or compound-shaped growth trajectories (Curran et al., 2010; Shanley, 2016; Thum & Kuhfeld, 2020), time-varying covariates (Shin et al., 2013), time-invariant covariates (Mok et al., 2015), and greater statistical power (Muthen & Curran, 1997). While researchers can use growth curve models to achieve many purposes, educational researchers commonly use these models to understand the mean academic or developmental status at a particular meaningful point in time (intercept) and mean growth rate over time (slope) (McNeish & Matta, 2017; Raudenbush & Bryk, 2002). Additionally, researchers can use growth curve models to examine variation around the mean (O'Connell & McCoach, 2008), and the influence of time-invariant (Mok et al., 2015; Shin et al., 2013) and time-varying covariates (Shin et al., 2013) to understand the ways in which estimated growth trajectories differ by contextual factors and child-level characteristics.

While there are many approaches to growth curve modeling, educational researchers frequently use one of two approaches, including latent growth curve modeling or multilevel modeling (MLM). The first approach, latent growth curve modeling, can be fitted under the structural equation modeling framework (Curran et al., 2010). Within this framework, researchers specify measurement models, impose a mean factor structure, and fix and free specific model parameters to examine different aspects of the growth process (McNeish & Matta, 2017). The second approach, MLM, can be conceptualized as an extension of the traditional linear regression model (Bryk & Raudenbush, 1987; O'Connell McCoach, 2008). Although the approach was originally developed to accommodate nested data structures, such as, children within classrooms or patients within a provider, the approach has been extended to include repeated measurements within individuals (Curran et al., 2010; Raudenbush & Bryk, 2002; Singer & Willet, 2003). Using this approach, researchers can nest assessment occasions within individuals and examine fixed effects such as mean growth rate or typical preliminary performance and random effects, such as the variability around the mean growth trajectory (McNeish & Matta, 2017).

Although researchers have repeatedly demonstrated that latent growth curves (LGCs) and MLMs are mathematically equivalent under many conditions (Bollen & Curran, 2006; Curran, 2003; Ferrer et al., 2009; Mehta & West, 2000), some researchers argue that given certain data structures, model specifications, and software capacities, researchers should choose one approach over the other (McNeish & Matta, 2017; O'Connell & McCoach, 2008; Curran et al., 2010). Researchers generally agree that MLMs should be used when researchers are modeling, a) three or more levels or b) data that are time-unstructured (O'Connell & McCoach, 2008; McNeish & Matta, 2017). Conversely, researchers generally suggest using LGCs when a) modeling data from measures that yield scores that are less reliable (Cole & Preacher, 2013), b) mediating or moderating relationships are of primary interest (O'Connell & McCoach, 2008), c) the nature of growth, e.g., shape of the growth trajectory is unknown (O'Connell & McCoach, 2008; McNeish & Matta, 2017) or d) specifying multigroup models with partial constraints (McNeish & Matta, 2017).

Using Growth Curve Models to Establish Norms. Although growth curve models are frequently used to model educational data and understand longitudinal trends in academic achievement and developmental progress (Shin et al., 2012; Mok et al., 2014; Shanley, 2016), few researchers have used GCMs to establish normative scores (Pan & Goldstein, 1997; Thum &

Kuhfeld, 2020). Pan and Goldstein (1997) used MLM to develop growth norms for young children that were conditional on previous height and weight measurements. Pan and Goldstein (1997) found that the novel method for establishing norms allowed for significant flexibility and easily captured the deceleration of growth over time. Similarly, Thum and Kuhfeld (2020) also viewed the methodology as an opportunity to obtain achievement and growth norms across content areas and grade levels conditioned on patterns of previous performance and instructional exposure. While the researchers of these two studies had different aims, they saw how the flexibility of the multilevel modeling methods could be leveraged to obtain normative scores that were conditional on variables of interest. While this method for deriving normative scores has not been widely adopted, both studies highlight the methodology as a viable option for establishing time-continuous normative scores.

Model fit. When researchers use LGCs and MLMs to understand longitudinal growth trends, researchers must examine model fit, or the use of selection criteria to choose between competing models, and model adequacy, or the ability of the independent variables to explain variability in the dependent variable (McCoach et al., 2022). Researchers using LGCs can use global fit indices, including the root mean square error of approximation (RMSEA) and the comparative fit index (CFI) to assess overall model fit (McNeish & Matta, 2017). Additionally, researchers can examine matrices of residual correlations, regression coefficients, and factor loadings to assess local fit and adequacy. However, researchers using MLMs have fewer options for assessing model fit and adequacy. Most of the strategies for assessing the fit of MLMs to the data include comparing two models and deriving information about relative fit, rather than absolute fit (McCoach et al., 2022). Researchers can use the likelihood ratio test (LRT) to compare nested models and the Akaike Information Criterion (AIC) and the Bayesian

Information Criterion (BIC) to compare non-nested models. Additionally, researchers can use pseudo-R^2 statistics to examine model adequacy. While there are many pseudo-R^2 statistics, educational researchers frequently use the proportional reduction in variance pseudo-R^2 statistic which represents the proportion of additional residual variance explained by the more parameterized model (Raudenbush & Bryk, 2002).

Modeling Time. Regardless of approach, researchers seeking to fit longitudinal data to growth curve models must determine how time will be measured and modeled. Singer and Willet (2003) acknowledged that there are many ways to measure the passing of time, but ultimately researchers should lean on theory and select the most appropriate time metric given their context and research aims. Educational researchers have modeled time in many ways, including, trimesters (Lambert, 2020), grade-level (Lambert, 2020; Mok et al., 2015; Shanley 2016; Thum & Kuhfeld, 2020), age in months (Hujar et al., 2021), and exposure to instruction (Thum & Kuhfeld, 2020). Generally, research has demonstrated that increasing the precision with which time is measured and modeled also supports goodness of model fit (Shanley, 2016; Thum & Kuhfeld, 2020). For example, Shanley (2016) used the ECLS-K 1998-1999 longitudinal dataset to estimate growth trajectories for K-8 students in mathematics using nine different LGC models. Time was modeled using grade-level and calendar year, the academic year, and the academic year with a summer discontinuity. Each model demonstrated increasingly better overall fit, suggesting that the specificity and accuracy of the model time metric significantly improved model fit. Similarly, in another study conducted by Thum and Kuhfeld (2020), researchers fitted longitudinal MAP data to MLMs to estimate intra- and inter-individual growth over time. To reduce random error, researchers used both grade level and approximate instructional exposure prior to assessment to create time-continuous normative ability and growth scores that were conditional on two different time metrics. Together these studies highlight the importance of choosing a significant and meaningful time metric to not only improve overall model fit, but also provide more accurate estimates of growth over time. More precise estimates are important for several key reasons, including that they, a) allow practitioners to understand how children are performing in comparison to individualized growth trajectories, b) facilitate data-driven goal setting (Thum & Kuhfeld, 2020), and c) allow policy makers to understand aggregate trends in achievement and establish necessary grants and legislation to facilitate better outcomes for subgroups of children (Kuhfeld et al., 2020).

Limitations and Directions for Future Research

Although researchers still use cross-sectional data analysis to estimate ability norms (Lambert, 2020), this method is unfavorable for several reasons. First, cross-sectional data does not adequately capture the intra-individual growth process (Raudenbush & Bryk, 2002; Singer & Willet, 2003; Thum & Kuhfeld, 2020). And second, researchers cannot rule out competing explanations for change over time (e.g., attrition, historical events) (Singer & Willet, 2003). Advances in growth curve modeling offer researchers a method for estimating ability norms based on intra-individual growth (O’Connell & McCoach, 2008; Thum & Kuhfeld, 2020) in the presence of missing data, time-unstructured data, and unbalanced data (Singer & Willet, 2003). Compared to previous methods which required all individuals to share the same assessment schedule and number of assessment occasions, growth curve modeling methods are far more flexible and allow researchers to maintain most, if not all individuals in the data set (Curran et al., 2010). Given that Lambert (2020) used more traditional cross-sectional methods to estimate *GOLD*® normative scores, future research should seek to fit longitudinal assessment data to GCMs to establish more precise and time-continuous norms. Using GCMs, assessment occasions could be nested within individual children and between-child variability could be leveraged to increase precision in normative estimates (Thum & Kuhfeld, 2020). By creating more precise norms conditioned on age and instruction would allow teachers to make more meaningful and accurate interpretations about relative performance.

Additionally, researchers have long cautioned against two-time-point data analysis in psychological and educational research due to measurement error in pre- and posttest scores (Curran et al., 2010; Lord, 1956) and the inability of a single difference score to capture change as a process (Singer & Willet, 2003). Yet, two-time-point data analysis methods are still frequently used to derive estimates of growth over time (Lambert, 2020). Given that Lambert (2020) calculated growth scores for each age-band and domain of learning and development by taking the difference between the average fall and spring scaled score, future research should use growth curve models to avoid the limitations of two-time-point data analysis methods and effectively demonstrate change as a process that occurs incrementally over time. Furthermore, using growth curve models to produce time-continuous normative scores would provide practitioners with a resource to understand typical growth between any two points in time, e.g., from zero to three months of instruction or from 60 months of age to 70 months of age.

Finally, while many researchers have acknowledged that LGCs and MLMs can yield nearly identical parameter estimates under many conditions (Bollen & Curran, 2006; Curran, 2003; Ferrer et al., 2009; Mehta et al., 2000), some researchers argue that given certain data structures, model specifications, and software capacities, researchers should select one approach over the other (McNeish & Matta, 2017; O’Connell & McCoach, 2008; Curran et al., 2010). For example, researchers suggest that in the presence of time-unstructured data, multilevel modeling should be considered (McNeish & Matta, 2017; O’Connell & McCoach 2008). While not impossible, researchers using LGCs to model time-unstructured data face significant challenges when preparing data for analysis. SEM software requires data to be processed in the wide or multivariate format, meaning that each time point requires a different column in the data set (McNeish & Matta, 2017). To combat this problem, researchers have collapsed time points into coarser categories and effectively created time groups. While this strategy eases the data processing burden, researchers lose valuable time information and potentially bias parameter estimates. Conversely, within the MLM framework, X is the time variable, therefore each individual can have a unique value for X. Furthermore, since MLM software processes data in the long or univariate format, each row includes a measurement occasion and corresponding time value. To avoid bias estimates and leverage between-child variation in time, e.g., age in months and instructional exposure, MLMs should be used to estimate time-continuous normative scores (Thum & Kuhfeld, 2020; McNeish & Matta, 2017).

CHAPTER THREE: METHOD

In Chapter Two, a thorough review of literature on the topics of *early childhood assessment*, *developmental growth trends*, and *approaches to modeling growth over time* set the stage for the current study which sought to model typical academic ability and developmental growth from birth through kindergarten using *GOLD*® domain-level scaled scores. While many of the themes and findings presented in Chapter Two are used to develop study hypotheses in Chapter Four and discuss the ways in which the results from the present study converge and diverge from previous findings in Chapter Five, Chapter Three focuses on the research methods and expands upon one approach to modeling change over time with time-unbalanced, time-unstructured, and nested assessment data. The sequence of topics for the present chapter include: the purpose of this study and corresponding research questions, an orientation to the measure, *GOLD*®, discussion of data collection and analysis procedures, methodological limitations, and ethical considerations.

Purpose

Current *GOLD*® normative ability and growth estimates lack precision for several reasons, including, a) the use of cross-sectional data to derive ability and growth normative scores, b) the use of two-time-point data to calculate gain scores, and c) variability in assessment schedules and children's ages at the time of assessment were treated as sources of random error. The present study seeks to provide more precise and time-continuous normative ability and growth scores for the average child at each point in time from birth through kindergarten using a series of multilevel models (Thum & Kuhfeld, 2020; Pan & Goldstein, 1997). To achieve the primary study aim and estimate the average child's preliminary developmental status and linear rate of growth for each age-band and domain of learning and development, two-level models were specified. Growth trajectories were modeled as a function of two different time metrics, including approximate months of instructional exposure and age in months. Model-based estimates were used to construct norm tables for each domain of development, age band, and time metric. In addition to developing time-continuous normative scores, additional study purposes included: a) provide evidence in support of the most effective time metric for modeling young children's developmental growth from birth through kindergarten across domains of learning and development b) provide evidence of the shape of the developmental pathway from birth through kindergarten for each domain of learning and development, and c) examine differences in preliminary status and linear growth rate over time by child-level characteristics. To achieve secondary study purposes, Akaike Information Criteria (AIC) tests and *Pseudo-*r^2 statistics were used to understand which time metric was the better predictor of academic and developmental growth (O'Connell & McCoach, 2008; McCoach et al., 2022). Next, linear slopes were examined sequentially from birth through kindergarten for each domain of learning or development to make inferences about the nature of the growth process. And finally, growth trajectories for different subgroups of children were examined by modeling child-level characteristics, including, gender, race/ethnicity, and primary language, at level-two, and examining resulting beta coefficients and corresponding *p*-values (O'Connell & McCoach, 2008).

Research Questions

1. What do model-based estimates suggest the average child will do at each point in time from birth through kindergarten, across each of the six major domains of learning and development presented in the *GOLD*® assessment system?

2. Which time metric is most effective for modeling growth, approximate instructional exposure or age in months, from birth through kindergarten?
3. What do hierarchical linear model-estimated slopes from age-separated cohort data suggest about the shape of the developmental pathway from birth through kindergarten across domains of learning and development?
4. How do growth trends differ between subgroups of children (e.g., by race/ethnicity, gender, and primary language)?

GOLD®

Data for the current study includes assessment records for children who were assessed using Teaching Strategies *GOLD*® during the 2021-2022 academic year. *GOLD*® is an observation-based authentic formative assessment designed to understand academic abilities and developmental growth in children ages birth through third grade (Burts et al*,* 2016), including children with disabilities and ELLs (Kim et al., 2013; Lambert, 2022). While *GOLD*® was designed to measure developmental and academic growth through third grade, the assessment is primarily used within early childhood settings, where higher ratings along developmental progressions are used to alleviate ceiling effects (Burts et al., 2016). The assessment system is comprised of four developmental domains including, social-emotional, physical, language, and cognitive, and five content domains including, mathematics, literacy, science and technology, social studies, and the arts. *GOLD*® also features a tenth domain, English language acquisition, for use with ELLs (Lambert, 2020). While the entire system features 10 domains, the current study sought to model typical developmental growth across the six major domains of learning and development identified by Teaching Strategies which include, literacy, mathematics, social-emotional, physical, cognitive, and language.

Assessment Purposes

According to Burts et al. (2016), the primary purposes of *GOLD*® include to help teachers, a) observe and document children's learning, b) support, guide, and aid instruction, c) identify children who may need further help or evaluation, and d) communicate with families and other stakeholders. Additionally, secondary purposes include a) collecting child performance data as part of a larger system of accountability and b) providing administrators with information to guide professional development opportunities for teachers. To support both primary and secondary assessment purposes, *GOLD*® was developed and has been externally validated for use as a formative, authentic, developmental, and criterion-referenced assessment (Lambert, 2017; Lambert, 2020). Recently, researchers have also provided validity evidence in support of norm-referenced score interpretations (Lambert, 2020).

Constructs

Teaching Strategies® hypothesized that each domain represented a unitary construct, and described the six latent constructs as follows:

> **Literacy:** Literacy ability begins at birth and progresses throughout early childhood, with beginning literacy reflecting "emergent reading and writing behaviors… [such as] verbal abilities, phonological sensitivity, familiarity with the basic purposes and mechanisms of reading, and letter knowledge (2016, p. 83). Furthermore, as children become readers, "they learn to decode unknown words, read with fluency, comprehend various types of text, and read for specific purposes and pleasure" (Burts et al., 2016, p. 84).
>
> **Mathematics:** Mathematics ability also begins from infancy and includes "mathematical vocabulary, concepts, and essential process skills... [such as] problem solving, reasoning, communicating, making connections, and representing" (Burts et al., 2016, p.111).

Language Development: "Learning to understand and use words. Language also involves learning about the structure and sequence of speech sounds, vocabulary, grammar, and the rules for engaging in appropriate and effective conversation (Burts et al., 2016, p. 41).

Social-Emotional Development: "Young children's social-emotional development involves learning how to understand their own and others' feelings, regulate and express their emotions appropriately, build relationships with others, and interact in groups" (Burts et al., 2016, p. 1).

Physical Development: "Physical development includes children's gross-motor (large muscle) and fine-motor (small muscle) skills. Balance; coordination; and locomotion, or traveling are part of the gross-motor development" (Burts et al., 2016, p. 23).

Cognitive Development: "Cognitive development, also called intellectual development, is influenced by the child's approaches to learning as well as his biological makeup and the environment" (Burts et al., 2016, p. 59). Furthermore, as children progress "they become more flexible and multidimensional in their thinking, solve a wider range of problems, mentally and symbolically manipulate concrete concepts, and think about their own mental activities" (Burts et al., 2016, p. 59).

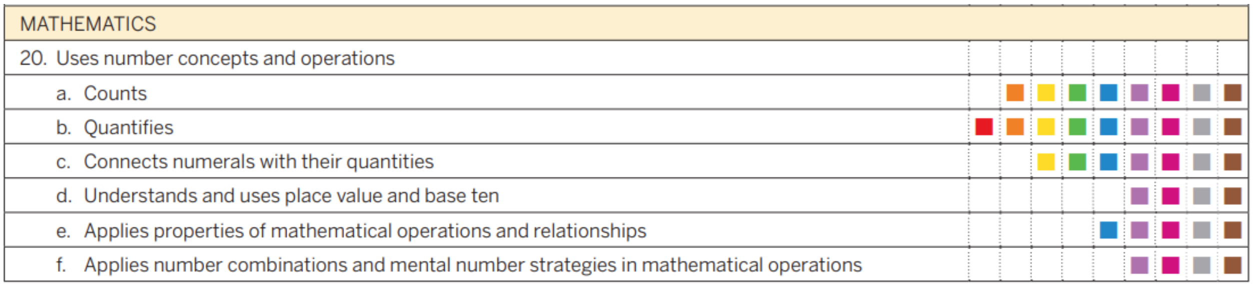
Dimensionality

Over time researchers have affirmed and reaffirmed scale dimensionality using multiple statistical methods. For example, Lambert et al. (2015) examined the plausibility of a six-factor model using a nationally representative sample of children's assessment records. Confirmatory factor analysis (CFA) results suggested the six-factor model fit the data reasonably well as evidenced by the following global fit statistics, Standardized Root Mean Squared Residual

(SRMR) =0.033, Comparative Fit Index (CFI) =.932, and Root Mean Square Error of Approximation (RMSEA) = 0.066. Local fit was also supported as evidenced by relatively large factor loadings (0.737-0.932). Lambert (2020) also tested the assumption of unidimensionality for each of the six major domains of learning and development using Principal Components Analysis of Rasch Residuals (PCAR). For each domain, the PCAR demonstrated that the Rasch dimension explained at least 90% of the variance in the data, thus providing further evidence in support of unidimensional constructs.

Assessment Structure

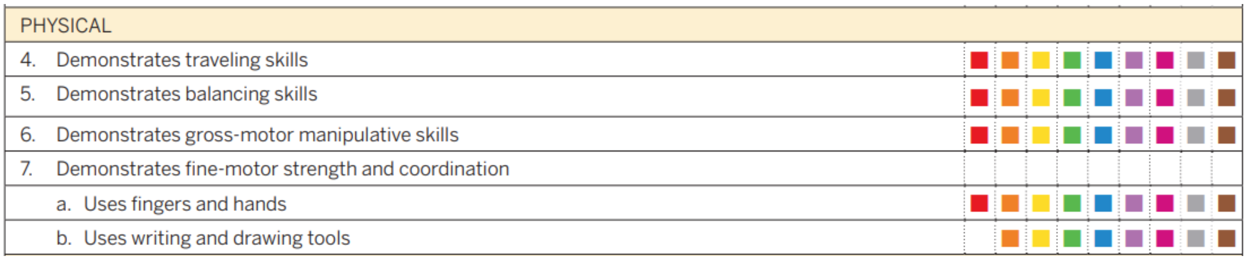
Within *GOLD*®, each domain of learning or development encompasses a set of objectives designed to guide teachers through the assessment process. Many objectives are further broken down into dimensions to better understand how children are progressing toward specific developmental and learning goals. For example, mathematics objective 20, *uses number concepts and operations,* features six unique dimensions, including, 20a) *counts,* 20b) *quantifies,* 20c) *connects numerals with their quantities,* 20d) *understands and uses place value and base ten*, 20e) *applies properties of mathematical operations and relationships*, and 20f) *applies number combinations and mental number strategies in mathematical operations.* For this objective and series of dimensions, teachers would document children's knowledge, skills, and abilities toward the six dimensions. Children would receive ratings for each dimension, but not the overarching objective. See Figure 1 for a visual.



MATHEMATICS									
20. Uses number concepts and operations									
a. Counts		🟧	🟨	🟩	🟦	🟪	◼	▨	🟫
b. Quantifies	🟥	🟧	🟨	🟩	🟦	🟪	◼	▨	🟫
c. Connects numerals with their quantities			🟨	🟩	🟦	🟪	◼	▨	🟫
d. Understands and uses place value and base ten						🟪	◼	▨	🟫
e. Applies properties of mathematical operations and relationships					🟦	🟪	◼	▨	🟫
f. Applies number combinations and mental number strategies in mathematical operations						🟪	◼	▨	🟫

Figure 1
Mathematics Objective 20 and Corresponding Dimensions a-f
From *GOLD® Objectives for Learning and Development; Birth Through Third Grade,* by Burts et al., 2016, p. xxvi. Image included with permission of Teaching Strategies, LLC. All rights reserved.

However, if objectives are not further broken down into unique dimensions, children receive ratings at the objective level. For example, children would receive ratings for objective four; *demonstrates traveling skills*, objective five; *demonstrates balancing skills*, and objective six; *demonstrates gross-motor manipulation skills.* See Figure 2 for a visual.

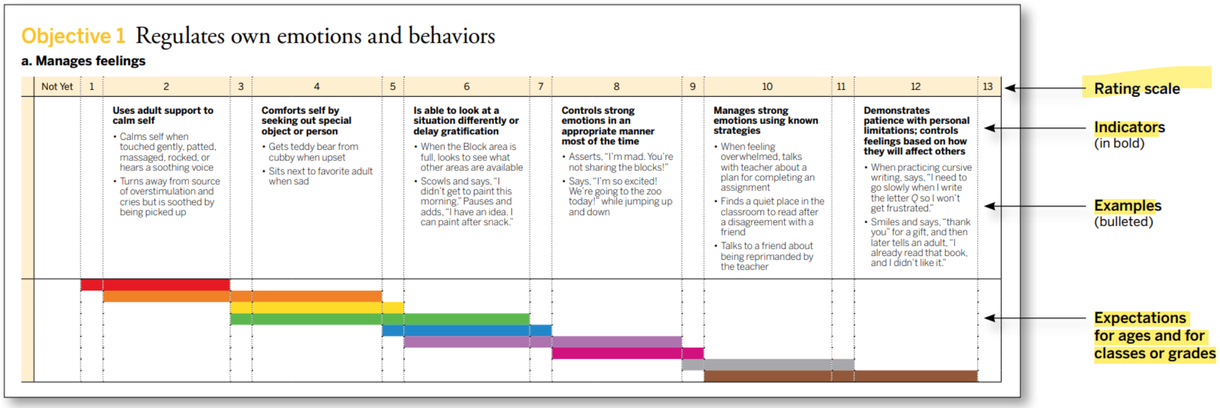


PHYSICAL									
4. Demonstrates traveling skills	🟥	🟧	🟨	🟩	🟦	🟪	◼	▨	🟫
5. Demonstrates balancing skills	🟥	🟧	🟨	🟩	🟦	🟪	◼	▨	🟫
6. Demonstrates gross-motor manipulative skills	🟥	🟧	🟨	🟩	🟦	🟪	◼	▨	🟫
7. Demonstrates fine-motor strength and coordination									
a. Uses fingers and hands	🟥	🟧	🟨	🟩	🟦	🟪	◼	▨	🟫
b. Uses writing and drawing tools		🟧	🟨	🟩	🟦	🟪	◼	▨	🟫

Figure 2
Physical Objectives 4- 7 and Dimensions 7a and 7b
From *GOLD® Objectives for Learning and Development; Birth Through Third Grade,* by Burts et al., 2016, p. xxiv. Image included with permission of Teaching Strategies, LLC. All rights reserved.

Progressions. The six major domains of learning and development feature developmental progressions. Progressions demonstrate the theoretical developmental pathway from birth through third grade for each unique objective or dimension. Each progression provides a sequence of expected knowledge, skills, and behaviors that children typically acquire on their journey toward mastery. Teachers can use progressions to assist in assigning ratings to child evidence, setting learning goals, planning instructional next steps, and communicating with families and other stakeholders (Lambert, 2020).

Each progression features an ordinal *rating scale*, ranging from 0 (not yet) to nine, 11, 13, 15, or 19, depending on the objective or dimension. Progressions also feature *indicators* and *examples* for each even rating along the expected pathway to aid teachers in understanding children's skills, abilities, and behaviors as they relate to particular steps. Finally, progressions include *color bands* or *expectations for ages and for classes or grades* that highlight where we would expect typically developing children to fall within a given year of life. The red color band represents expected development for children ages birth to one-year-old, orange is for children ages one- to two-years old, yellow is for children ages two- to three-years old, green is for children in preschool, blue is for children in pre-k, purple is for children in kindergarten, and pink, gray, and brown are for children in first, second, and third grade respectively. See Figure 3 for an example.



Objective 1 Regulates own emotions and behaviors

a. Manages feelings

Not Yet	1	2	3	4	5	6	7	8	9	10	11	12	13
		Uses adult support to calm self • Calms self when touched gently, patted, massaged, rocked, or hears a soothing voice • Turns away from source of overstimulation and cries but is soothed by being picked up		**Comforts self by seeking out special object or person** • Gets teddy bear from cubby when upset • Sits next to favorite adult when sad		**Is able to look at a situation differently or delay gratification** • When the Block area is full, looks to see what other areas are available • Scowls and says, "I didn't get to paint this morning." Pauses and adds, "I have an idea. I can paint after snack."		**Controls strong emotions in an appropriate manner most of the time** • Asserts, "I'm mad. You're not sharing the blocks!" • Says, "I'm so excited! We're going to the zoo today!" while jumping up and down		**Manages strong emotions using known strategies** • When feeling overwhelmed, talks with teacher about a plan for completing an assignment • Finds a quiet place in the classroom to read after a disagreement with a friend • Talks to a friend about being reprimanded by the teacher		**Demonstrates patience with personal limitations; controls feelings based on how they will affect others** • When practicing cursive writing, says, "I need to go slowly when I write the letter Q so I won't get frustrated." • Smiles and says, "thank you" for a gift, and then later tells an adult, "I already read that book, and I didn't like it."	

Figure 3

Progression for Objective 1a: Regulates Own Emotions and Behaviors

From *GOLD*$^{®}$ *Objectives for Learning and Development; Birth Through Third Grade,* by Burts et al., 2016, p. x. Image included with permission of Teaching Strategies, LLC. All rights reserved.

Scores

In the most recent *GOLD*® technical manual, evidence is provided to support four distinct score interpretations, including raw scores, widely held expectation (WHE) scores, scaled scores, and national normative scores (Lambert, 2020). Each score interpretation is explained in greater detail below:

Raw Scores: Raw scores are calculated when all finalized ratings within a given domain are summed. Raw scores represent an ordinal level of measurement and reflect a child's current skills and abilities related to the corresponding domain of learning or development (Lambert, 2020). Since raw scores are a summation of all ratings on all progressions corresponding to one domain, raw scores can range significantly from one domain to another. For example, in literacy, a domain with five objectives and 16 dimensions, children can obtain raw scores ranging from zero to nearly 200 points. Yet, in physical development, a domain with four objectives and only two dimensions, children can obtain raw scores ranging from zero to nearly 70 points. Furthermore, since some progressions feature ordinal scales ranging from zero to nine points, while others feature scales ranging from zero to 19 points, some items have greater influence over the domain-level raw score than other items.

WHE Scores: Raw scores are also used to determine if children are performing below, within, or beyond the WHE range for each domain of learning or development. The WHE range can be determined by summing the lower WHE bound for each progression within a given domain and the upper WHE bound for each progression with a given domain. If a child's raw score for the given domain falls between the lower and upper WHE bounds for the given domain, the child is meeting expectations. Conversely, if the child's score falls below the lower bound of the WHE range, the child is performing below expectations and if the child's score resides above the upper bound of the WHE range then the child is exceeding expectations. While WHE scores also reflect an ordinal level of measurement, they can be used to understand how a child's present level of performance within a given domain compares to WHEs for children within the same year of life (Lambert, 2020).

Domain-Level Scaled Scores: Scaled scores are obtained by transforming raw scores using a Rasch measurement model. Unlike raw scores and WHE scores which reflect an ordinal level of measurement, scaled scores have interval-level properties and allow educators and researchers to understand developmental and academic growth over time. Additionally, since scaled scores reside on a single scale ranging from 0-1000, children's domain-level scaled scores can be compared meaningfully even if children are at different developmental levels (Lambert, 2020). Finally, because scaled scores account for missing values, they provide a reasonable ability estimate in the presence of missing values at the item-level.

National Normative Scores: National normative scores were obtained using nationally representative samples of children's domain-level scaled scores. For each trimester, year of life, and domain, scaled score distributions were divided into quartiles. Performance at the 25th, 50th, and 75th percentiles were reported. Teachers can use national normative ability scores to understand how their children are performing in comparison to a nationally representative sample of children at particular points throughout the academic year. Additionally, growth norms capture average change in domain-level scaled scores between the fall and spring measurement occasions for each domain and given year of life. Teachers can use growth estimates to understand average growth for each year of life (Lambert, 2020).

To aid in meeting numerous assessment purposes, *GOLD*® yields four unique scores. Educators may use raw scores or WHE scores to plan for future instruction, identify children who may need additional support, and communicate with families. However, scores expressed using ordinal levels of measurement, such as raw scores or WHE scores, do not allow for understanding growth across trimesters or years. Educators and administrators may use interval-level scaled scores to understand how much growth a child made from the fall of pre-k to the spring of pre-k or to understand if a child made more growth during their preschool or pre-k year. Furthermore, because scores are placed on a common scale, meaningful comparisons can be made between children who are at different developmental levels (Lambert, 2020).

Reliability. Since the present study sought to understand typical development from birth through kindergarten, domain-level scaled scores were used so that preliminary developmental abilities and growth rates could be meaningfully compared over time and for children who were at different developmental levels. Lambert (2020) acknowledged that "reliability and validity are not inherent qualities of an assessment but rather are properties of the information an assessment provides under particular conditions of use" (p. 4). Given that evidence of reliability is prerequisite to valid score interpretations, Lambert (2020) provided information regarding person-separation reliability, item-separation reliability, and internal-consistency reliability for *GOLD*® domain-level scaled scores. High person-separation reliability coefficients suggest that there is a high probability of replicating the same separation of persons across multiple measurements. For each domain of learning or development, person-separation reliability coefficients were greater than or equal to .96. Similarly, high item-separation reliability coefficients suggest that it's very likely to obtain the same separation of items across multiple measurements. Lambert (2020) found that item-separation reliability coefficients were greater than or equal to 0.99 across domains of learning and development. And finally, internal consistency reliability as measured by coefficient alpha was greater than or equal to .97 for each domain of learning or development. In summary, there is substantial evidence to support the reliability of *GOLD*® scaled scores for children ages birth through first grade across the six major domains of learning and development (Lambert, 2020).

Measurement Invariance. While evidence of scaled score reliability is one important source of validity evidence, another critical source includes evidence of measurement invariance. According to The American Educational Research Association et al. (2014),

> A test that is fair within the meaning of *The Standards* reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population: a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct (p. 50).

To provide evidence that constructs were interpreted similarly across children who belong to different subgroups, researchers used multigroup CFA to examine measurement invariance (Lambert, 2022). Resulting fit statistics supported strict measurement invariance between boys and girls (CFI=.997, TFI=.997, RMSEA= .049, 95% CI[.049,.049], SRMR=.044), children whose primary language was English and children whose primary language was Spanish (CFI=.996, TFI=.996, RMSEA=.057 95% CI[.057,.057], SRMR=.050], and children who identify as White and non-White (CFI=.997, TFI=.997, RMSEA=.053 95% CI[.053,.054], SRMR=.051). Given evidence of strict invariance, meaningful comparisons can be made across latent factor means for boys and girls, White and non-White children, and children whose primary language is English and children whose primary language is Spanish (Pirralha, 2020).

Data Collection Procedures

Teachers engage in the assessment process as they collect evidence of children's learning and development during routine activities, compile evidence into portfolios, make preliminary placements along developmental and learning progressions, and finalize ratings at predetermined timepoints throughout the academic year (Lambert, 2020). Many teachers finalize ratings three times throughout the year, including once at the end of each trimester. However, frequency of use varies from program-to-program and state-to-state. When teachers finalize ratings along developmental and learning progressions, item-level data are entered into the Teaching Strategies *GOLD*® online platform. Item-level scores are summed to obtain domain-level raw scores. Next, domain-level raw scores are transformed to scaled scores using a Rasch measurement model. The model uses a proprietary algorithm to impute item-level data for any child that has received ratings on at least 80% of progressions within each domain. The model is based on mean substitution and uses other ratings within the domain to estimate missing values.

Variables

The 2021-2022 dataset included item-level ratings, domain-level raw scores, and domain-level scaled scores. In addition to assessment data, most children had demographic information on record, including, grade-level, birthdate, gender, race, ethnicity, and primary language and some children had information on record, including, disability status and FRL-eligibility status. Several additional variables were created and added to the 2021-2022 data set. Model variables and demographic variables are described in greater detail below.

Social Emotional, Language, Physical, and Cognitive Domain-Level Scaled Scores: calculated by transforming the social-emotional, language, physical, and cognitive domain-level raw scores using a Rasch measurement model. The resulting scores represent developmental status on the date the ratings were finalized for each respective construct. Scores reflect an interval-level of measurement and reside on a scale that ranges from 0-1000.

Literacy and Mathematics Domain-Level Scaled Scores: calculated by transforming the literacy and mathematics domain-level raw scores using a Rasch measurement model. The resulting scores represent academic ability on the date the ratings were finalized for each respective construct. Scores reflect an interval-level of measurement and reside on a scale that ranges from 0-1000.

Checkpoint Finalization Date: This is a date variable and reflects the date a child's ratings were finalized in the online platform for a particular checkpoint, e.g., the fall social-emotional checkpoint or the winter mathematics checkpoint.

Approximate Instructional Exposure: This variable was created by taking the assessment finalization date for each domain and measurement and subtracting the typical school start date, August 14, 2021. This variable represents approximate instructional exposure in months and ranges from 0 to 11.

Age in Months: This variable was created by taking the assessment finalization date for each domain and measurement and subtracting the child's birthdate. This variable represents each child's age at the time each assessment score was finalized. This variable ranges from 0 to 80.

Age in Months Centered: This variable was created by taking each child's age in months at the time of assessment and subtracting the first age included in each age band. This variable represents a child's additional months of age from the first age included in the age band at the time of each assessment finalization date. This variable ranges from 0-20.

Birthdate: This is a date variable and reflects the date a child was born. The birthdate variable was used to calculate the child's age in months at each desired point in time (e.g., beginning of the academic year).

Age-Band: This categorical variable includes six distinct categories, birth to one-year-old children, one- to two-year-old children, two- to three-year-old children, preschool children (three- to four-year-olds), pre-k children (four- to five-year-olds), and kindergarten children (five- to six-year-olds). While each child is assigned to an age-band based on their classroom level, it is possible to have children within each color band that reside outside of the intended age range served. For example, it's likely that there are some seven-year-old children in kindergarten classrooms.

Academic Year Start Date: This date variable was established using the average start date for the academic year in the U.S., August 14, 2021, and is the same for all children in the sample (DeSilver, 2019).

Age Cut Date: This date variable was created using a typical cut date to determine eligibility to start kindergarten. This date is September 1, 2021, and is the same for all children in the sample (O'Connor, 2019).

Gender: This is a dichotomous categorical variable; children can be identified as male (0) or female (1).

Race/Ethnicity: This categorial variable combines race and ethnicity and features the same categories as the U.S. Census, including, White Non-Hispanic, Black Non-Hispanic, Asian Non-Hispanic, Native American, Native Hawaiian, Alaskan Native, or Pacific Islander Non-Hispanic, Bi- or Multiracial Non-Hispanic, and Hispanic or Latino. Dummy variables were created for each race/ethnicity category.

Primary Language: This categorical variable represents a child's first language and includes the following categories: English, Spanish, and Other. Dummy variables were created for each language category.

Free- and Reduced-Price Lunch Eligibility Status: This is a dichotomous categorical variable; children can be identified as Free- and Reduced-Price Lunch eligible (1) on ineligible (0).

Individualized Family Service Plan Eligibility Status: This is a dichotomous categorical variable; children can have an Individualized Family Service Plan (1) or not (1).

Individualized Education Plan Eligibility Status: This is a dichotomous categorical variable; children can have an Individualized Education Plan (1) or not (1).

Population

During the 2021-2022 academic year, nearly 1.3 million children in birth through kindergarten classrooms were assessed using the English version of *GOLD*®, including children from all 50 states and Washington, D.C. Results from the current study seek to provide normative ability and growth scores that can be generalized to children who are in birth to one-year-old, one- to two-year-old, two- to three-year-old, preschool, pre-k, and kindergarten classrooms and are assessed using *GOLD*®.

Eligibility Criteria

Prior to sampling, several criteria were established to support the development of precise time-continuous normative scores and eliminate some degree of construct-irrelevant variance. First, while there are English and Spanish versions of *GOLD*®, only children assessed using the English-version of *GOLD*® were considered for the current study, as researchers have not established measurement invariance between the Spanish- and English-versions of *GOLD*®. Next, the number of children nested within each rater was reviewed. To capture typical classroom use, the National Association for the Education of Young Children's *Staff-to-Child Ratio and Class Size* guidelines (2018) and empirical data were considered. Children in birth to one year old classrooms were eligible to be sampled if they shared a rater with one to 19 other children, children in one- to two-year-old and two- to three-year-old classrooms were eligible if they shared a rater with three to 19 other children, and preschool, pre-k, and kindergarten children were eligible to be sampled if they shared a rater with five to 24 other children. This criterion served two purposes, including a) to capture typical use which includes teachers assigning ratings to children in their own classrooms and b) to reduce nesting effects which may be induced if a single rater assigned ratings to all the children in one childcare center. Additionally, to enact the sampling plan, age band and race/ethnicity data were needed. Therefore, only children with race, ethnicity, and age band data were eligible to be sampled. Finally, to be considered for the current study, children had to be reasonably aged for their age band/grade level. The lower age threshold for each age band or grade was the first age included in the color band (e.g., children in birth to one-year-old classrooms had to be zero months as of September 1, 2021, and children in kindergarten classrooms had to be at least 60 months on September 1, 2021). The upper threshold for each age band was determined by examining empirical distributions. This criterion was implemented to support the development of normative scores that were not overly influenced by outliers and reflective of typical development at each point in time from birth through kindergarten. See Table A1 in Appendix A for more information about children's ages at the age cut date, September 1, 2021.

Sampling Strategy

Stratified random sampling was used to derive a nationally representative sample of children ages birth through kindergarten who were assessed using *GOLD*® during the 2021-2022 academic year. First, the sampling frame was divided into six strata using children's age/grade bands, including, birth to one-year-old children, one- to two-year-old children, two- to three-year-old children, preschool children (three- to four-year-olds), pre-k children (four- to five-year-olds), and kindergarten children (five- to six-year-olds). Next, 2020 Census race/ethnicity data for children was used to determine the proportion of children to sample from each stratum. According to the 2020 Census data, children who identified as White Non-Hispanic made up about one-half of the population of children in the U.S. (*N*=36,854,828, 49.65%). Children who identified as Hispanic or Latino made up approximately one-fourth of the population (*N*=18,983,948, 25.58%). Children who identified as Black Non-Hispanic made up nearly one-seventh of the population (*N*=10,196,603, 13.74%). And finally, children who identified as Asian Non-Hispanic (*N*=4,011,508, 5.40%), two or more races Non-Hispanic (*N*=3,412,289, 4.60%), or Native Hawaiian Non-Hispanic, Pacific Islander Non-Hispanic, Alaskan Native Non-Hispanic, and American Indian Non-Hispanic (*N*=762,457, 1.03%) accounted for small proportions of the population (Population Division U.S. Census Bureau, 2022, as cited by Annie E. Casey Foundation, 2023). Given a target sample of 3,000 children per age-band stratum, the target sample per age band strata included 1,490 White Non-Hispanic children, 767 Hispanic or Latino children, 412 Black Non-Hispanic children, 162 Asian Non-Hispanic children, 138 Bi- or Multiracial Non-Hispanic children, and 31 Native Hawaiian Non-Hispanic, Pacific Islander Non-Hispanic, Alaskan Native Non-Hispanic, or American Indian Non-Hispanic children.

Sample

The full sample included assessment records from 18,000 children who resided in classrooms where *GOLD*® checkpoints were finalized three times during the 2021-2022 academic year. Children ranged in age from zero to 70 months on September 1, 2021 (*M* =34.33, *SD*=21.04). The sample was split relatively evenly between boys (*n*=9,330, 51.88%) and girls (*n*=8,655, 48.12%). The race/ethnicity breakdown closely mirrored the Census records and was as follows: White Non-Hispanic (*n*=8,983, 49.91%) Hispanic or Latino (*n*=4,624, 25.69%), Black Non-Hispanic (*n*=2,484, 13.80%), Asian Non-Hispanic (*n*=890, 4.94%), Bi- or Multiracial Non-Hispanic (*n*=832, 4.62%), and Native Hawaiian Non-Hispanic, Pacific Islander Non-Hispanic, Alaskan Native Non-Hispanic, and American Indian Non-Hispanic (*n*=187, 1.04%). Asian children were slightly underrepresented in the birth- to one-year-old (*n*=134) and the kindergarten (*n*=140) sample. Most children's primary language was English (*n*=13,429, 74.61%), followed by Spanish (*n*=2,454, 13.63%), and other (*n*=2,117, 11.76%). Nearly one-third of children sampled (*n*=4,993, 27.74%) were eligible for free- or reduced-price lunch (FRL). A small proportion of children sampled were on an Individual Family Support Plan or Individualized Education Plan (*n*=1,226, 6.81%) and a small proportion were identified as English Language Learners (*n*=1,635, 9.08%). Additional demographic data and relevant descriptive statistics are provided in Table A2 in Appendix A.

Data Analysis Procedures

Longitudinal *GOLD*® data were fitted to a series of MLMs to obtain estimates of the fixed effects and establish time-continuous ability and growth normative scores for children ages birth through kindergarten across domains of learning and development. Compared to previous methods used to establish norms which included cross-sectional and two-time point data, MLMs were used to produce more precise estimates that were reflective of the intra-individual growth process and leveraged between-child variability in age and instructional exposure (Pan & Goldstein, 1997; Thum & Kuhfeld, 2020). Multilevel models were used, rather than LGCs, for two primary reasons, including researchers a) suggest that MLMs are preferable in the presence of time-unstructured data (Curran et al., 2010; McNeish & Matta, 2017; O'Connell & McCoach, 2008) and b) have used MLMs to establish normative ability and growth estimates in other contexts (Pan & Goldstein, 1997; Thum & Kuhfeld, 2020). Within the MLM framework, HLMs were selected because linear models typically fit the data reasonably well when individuals have three measurements that were collected over a relatively short time-period (Raudenbush & Bryk., 2002).

Measuring and Modeling Time

Researchers seeking to model developmental and academic growth using hierarchical linear growth models must select a time metric that is meaningful to their phenomenon and research aims. After reviewing numerous studies that sought to describe the growth process across domains of learning and development for children ages birth through kindergarten and beyond, it was apparent that researchers have used many different units to measure and model the passing of time. Researchers have used trimesters (Lambert, 2020), grade-level (Mok et al., 2015; Shanley, 2016; Thum & Kuhfeld, 2020), instructional exposure, (Thum & Kuhfeld, 2020) and age in months (Hujar et al., 2021). Yet, none of these studies sought to understand continuous growth trends from birth through kindergarten. Furthermore, while researchers seeking to model older student's academic growth often explored numerous approaches to measuring and modeling time (Shanley, 2016; Thum & Kuhfeld, 2020), literature was void of studies that sought to understand the most effective time metric for modeling young children's developmental growth which is frequently rapid (Center for the Developing Child, 2016) and closely related to age (Harkness et al., 2013).

Given the limitations of previous research, the current study modeled the passing of time using two different time metrics: age in months and approximate months of instructional exposure at the time of each assessment finalization date. Normative score estimates based on months of instructional exposure may be useful in understanding typical performance for children who are also typically aged and children in older age bands, while normative estimates based on months of age may be more appropriate for babies and toddlers where developmental progress is closely related to age (Harkness et al., 2013).

Centering Time. Once a time metric is selected, researchers must also choose a centering strategy that allows for meaningful interpretations of intercept parameters (O'Connell & McCoach, 2008; Raudenbush & Bryk, 2002). For each age band, age in months was centered on the first age included in the age band. For example, each model for babies in birth to one-year-old classrooms was centered on zero month. Likewise, each model for kindergarteners was centered on 60 months. Using this strategy, age intercepts can be interpreted as the expected domain-level scaled score for the youngest children in the age band. Additionally, across all age-bands, instructional exposure was centered on the average school start date, August 14, 2021

(DeSilver, 2019). Using this centering strategy, instructional exposure intercepts can effectively be interpreted as the expected domain-level scaled score for children prior to instruction.

Data Requirements and Assumptions

Researchers seeking to model growth using hierarchical linear growth models must examine several assumptions prior to conducting analyses. If assumptions are ignored and misspecification occurs at level one, level-one and level-two estimates can be biased (Raudenbush & Bryk, 2002). Individual age-band files were reviewed for each data requirement and assumption and results are discussed in subsequent paragraphs.

Missing data. While HLMs allow for missing data, data must be missing completely at random (MCAR) or missing at random (MAR) (O'Connell & McCoach, 2008; Raudenbush & Bryk, 2002). Data are considered MCAR if missing values are randomly distributed across the variable and unrelated to other variables, while data are considered MAR if missing values are not randomly distributed and are related to other variables in the dataset (Raudenbush & Bryk, 2002). Little's MCAR test (1988) was used to examine whether data were MCAR. Large Chi-Square statistics and statistically significant *p*-values suggested that the missing values were not MCAR across age bands. To examine the plausibility that data were MAR, missing values were carefully examined in relation to other observed values. Through careful review, a pattern of missingness emerged between site IDs and domain-level scaled scores, where the missingness of the domain-level scaled score could be explained by the child's site ID. Given that the propensity of missing values was determined to be related to other observations, the mechanism of missingness was determined to be MAR.

While researchers can use HLMs to model incomplete data, missing data should be handled carefully to ensure unbiased estimates (Raudenbush & Bryk, 2002). After careful examination, it was determined that data were missing for less than 7% of cases for each variable. Researchers have suggested that the benefit to multiple imputation is insignificant below 5% missingness and estimates are unlikely to be biased by missing data below 10% (Lee & Huber, 2021). Therefore, missing values were not imputed. Instead, Full Information Maximum Likelihood was used to obtain parameter estimates in the presence of partially missing data (Raudenbush & Bryk, 2002).

Observations and Parameter Estimates. First, hierarchical linear modeling requires that individuals have at least one measurement (Raudenbush & Bryk, 2002). This requirement was met through preliminary data cleaning procedures. Next, HLMs require one more wave of data than the number of growth parameters included in the level one model (O'Connell & McCoach, 2008). This data requirement was met given that most children had three measurements and only two level-one parameters were estimated (slope and intercept).

Linearity. HLM assumes a linear relationship between the independent and dependent variables (Palmeri, n.d.). To examine the tenability of this assumption, SPSS 28 was used to create scatter plots using children's age in months and model residuals and children's instructional exposure and model residuals. Across age bands, domains, and time metrics, plots demonstrated a random scatter, thus providing support for the tenability of this assumption.

Homogeneity of Variance. To examine the tenability of this assumption, SPSS 28 was used to create scatter plots of fitted values and residuals (Palmeri, n.d.). Across age bands, domains, and time metrics, plots demonstrated a random scatter, thus supporting the tenability of this assumption.

Normality of Residuals. HLM assumes normality of residuals (Palmeri, n.d.). To examine the tenability of this assumption, QQ plots were used to examine the relationship between the standardized residuals and normal quantiles. While there was some deviation at the tails, overall, the residuals demonstrated a linear diagonal pattern across age bands, domains, and time metrics, providing evidence to support the tenability of this assumption.

Software and Estimation Procedures

Hierarchical linear growth models were fitted to longitudinal data using HLM 8 software. To derive model-based estimates, full information maximum likelihood (ML) estimation procedures were used. Under ML, the software estimates values for the level-two fixed coefficients, γ, Tau matrix, T, and level-one residual variance, σ^2 that maximize the joint likelihood of all three parameters given the sample data, Y (Raudenbush & Bryk, 2002). ML was selected over the default estimation procedure, restricted information maximum likelihood (REML) for two reasons, including a) ML is preferable in the presence of missing data (Raudenbush et al., 2002) and b) ML is required to compare non-nested models using Information Criteria tests such as the AIC test (McCoach et al., 2022).

Modeling Typical Development

To answer the first research question, which sought to understand typical or average development at each point in time from birth through kindergarten across each of the six major domains of learning and development presented in the *GOLD*® assessment system, longitudinal assessment data were fitted to a series of MLMs (Pan & Goldstein, 1997; Thum & Kuhfeld, 2020). First, data were fitted to 36 completely unconditional models, one for each age band and domain of learning and development. Completely unconditional models were specified for three reasons, including to, a) decompose variance, b) calculate the intraclass correlations (ICCs), and c) obtain baseline model fit statistics. Variance components, including the proportion of between-child and within-child variance, were used in subsequent analyses and ICCs were examined to determine if a significant proportion of the total variance resided between children and therefore provided justification for nesting assessment occasions within children.

Equation 1 includes the Completely Unconditional Level I Model and was used to model a child's domain level scaled score, Y_{ti}, as a function of the child's average academic or developmental level across measurements, π_{0i}, and a residual term, e_{ti}. Equation 2 includes the Completely Unconditional Level II Model and was used to model a child's average academic or developmental level across measurements, π_{0i}, as a function of the grand mean academic or developmental level, β_{00}, and an error term, r_{0i}. Although the subscript *t* was used in the Completely Unconditional Model, time was not modeled in this set of equations, rather, π_{0i} is an estimate of each child's average academic or developmental level across measurements and β_{00} represents the grand mean academic or developmental level across children. While researchers typically begin with the Unconditional Model, rather than the Completely Unconditional Model, the Completely Unconditional Model was required for partitioning variance and more importantly, obtaining proportion of variance reduction statistics for more parameterized models, including the Unconditional Instructional Exposure Growth Model and the Unconditional Age in Months Growth Model. Furthermore, while the Completely Unconditional Level I and Level II Models are the same across time metrics, a second set of Completely Unconditional Level I and Level II Models with different subscripts are provided in Equations 3 and 4, so that subsequent, nested or more parameterized models can maintain the same set of subscripts.

Completely Unconditional Level I Model (Instruction)

$$Y_{ti} = \pi_{0i} + e_{ti} \tag{1}$$

Where:

Y_{ti} = academic or developmental level for child *i* at time *t*

π_{0i} = average academic or developmental level for child *i* across measurements

e_{ti} = within child residual

Completely Unconditional Level II Model (Instruction)

$$\pi_{0i} = \beta_{00} + r_{0i} \tag{2}$$

Where:

π_{0i} = average academic or developmental level for child *i* across measurements

β_{00} = grand mean academic or developmental level

r_{0i} = random error in measurement for child *i*

Completely Unconditional Level I Model (Age in Months)

$$Y_{ti} = \pi_{2i} + e_{1ti} \tag{3}$$

Where:

Y_{ti} = academic or developmental level for child *i* at time *t*

π_{2i} = average academic or developmental level for child *i* across measurements

e_{1ti} = within child residual

Completely Unconditional Level II Model (Age in Months)

$$\pi_{2i} = \beta_{20} + r_{1i} \tag{4}$$

Where:

π_{2i} = average academic or developmental level for child *i* across measurements

β_{20} = grand mean academic or developmental level

r_{1i} = random error in measurement for child i

Next, 72 unconditional hierarchical linear growth models were specified, one for each age band, domain of learning and development, and time metric. Intercepts were allowed to vary for all growth models, as evidenced by level-two residual terms, r_{1i} and r_{2i}. Model-estimated random intercept reliability coefficients suggested intercept values were generally reliable across Unconditional Instructional Exposure Growth Models (.76-.93) and Unconditional Age in Months Growth Models (.71-.94). While intercepts were allowed to vary between children, linear slopes were fixed. To examine whether random slope models demonstrated significantly better fit to the data, random slope coefficients were added to Kindergarten Level II Instructional Exposure Growth Models and Level II Age in Months Growth Models. Statistically significant chi-squared values resulting from Likelihood Ratio Tests (LRT) suggested that random-slope models demonstrated better fit to the data. However, LRTs which rely on chi-squared tests are notoriously sensitive to large sample sizes (McCoach et al., 2022). Therefore, model AICs were also reviewed. Minimal changes in AIC values across nested models suggested the additional error term was unnecessary and therefore, the more parsimonious models were retained.

Equation 5 includes the Unconditional Level I Instructional Exposure Growth Model and was used to model the domain-level scaled score for child i at time t, Y_{ti}, as a function of the child's academic ability or developmental level prior to instruction, π_{0i}, a linear rate of growth, π_{1i}, additional months of instruction since the beginning of the academic year, (*instruction – instruction*$_1$)$_{ti}$, and a residual term, e_{ti}. The Unconditional Level II Instructional Exposure Growth Model includes equations 6 and 7. Equation 6 was used to model each child's academic or developmental level prior to instruction, π_{0i}, as a function of the average academic ability or developmental level prior to instruction, β_{00}, and an error term, r_{0i}. Equation 7 was used to model each child's linear rate of growth, π_{1i}, as a function of the average growth rate, β_{10}.

Unconditional Level I Instructional Exposure Growth Model

$$Y_{ti} = \pi_{0i} + \pi_{1i}(instruction - instruction_1)_{ti} + e_{ti} \tag{5}$$

Where:

Y_{ti} = academic or developmental level for child *i* at time *t*

π_{0i} = academic or developmental level prior instruction for child *i*

π_{1i} = linear rate of growth for child *i*

instruction = months of instruction for child *i* at time *t*

$instruction_1$ = 0 months of instruction for child *i*

e_{ti} = within child residual at time *t*

Unconditional Level II Instructional Exposure Growth Model

$$\pi_{0i} = \beta_{00} + r_{0i} \tag{6}$$

$$\pi_{1i} = \beta_{10} \tag{7}$$

Where:

π_{0i} = academic or developmental level prior to instruction for child *i*

β_{00} = average academic or developmental level prior to instruction

r_{0i} = random error in intercept for child *i*

π_{1i} = linear rate of growth for child *i*

β_{10} = average linear growth rate

Equation 8 includes the Unconditional Level I Age in Months Growth Model and was used to model the domain-level scaled score for child i at time *t*, Y_{ti}, as a function of the child's academic ability or developmental level at the beginning of the age band, π_{2i}, a linear rate of growth, π_{3i}, additional months of age from the first age in that age band (age – *age*$_{1}$)$_{ti}$, and a within child residual term, e_{1ti}. The Unconditional Level II Age in Months Growth Model includes equations 9 and 10. Equation 9 was used to demonstrate each child's academic or developmental level at the beginning of the age band, π_{2i}, as a function of the average academic ability or developmental level at the beginning of the age band, β_{20}, and an error term, r_{1i}. Equation 9 was used to model each child's linear rate of growth, π_{3i}, as a function of the average growth rate, β_{30}.

Unconditional Level I Age in Months Growth Model

$$Y_{ti} = \pi_{2i} + \pi_{3i}(age - age_1)_{ti} + e_{1ti} \qquad (8)$$

Where:

Y_{ti} = academic or developmental level for child *i* at time *t*

π_{2i} = academic or developmental level for child *i* at the youngest age in the age band

π_{3i} = linear rate of growth for child *i*

age = age in months for child *i* at time *t*

age_1 = first age in age band, measured in months

e_{1ti} = within child residual at time *t*

Unconditional Level II Age in Months Growth Model

$$\pi_{2i} = \beta_{20} + r_{1i} \qquad (9)$$

$$\pi_{3i} = \beta_{30} \tag{10}$$

Where:

π_{2i} = academic or developmental level for child *i* at the youngest age in the age band

β_{20} = average academic or developmental level at the youngest age in the age band

r_{1i} = random error in intercept for child *i*

π_{3i} = linear rate of growth for child *i*

β_{30} = average linear growth rate

The Most Effective Approach to Modeling Time

To answer the second research question which sought to provide evidence of the most effective approach to measure and model time for children ages birth through kindergarten across domains of learning and development, two approaches were used. First, level-one *pseudo-r^2* statistics, including the proportional reduction in variance statistics, were calculated for each set of nested models (e.g., the Completely Unconditional Kindergarten Language Model and the Unconditional Kindergarten Age in Months Language Model) (Raudenbush & Bryk, 2002 as cited in McCoach et al., 2022). Next, *pseudo-r^2* values were compared for each set of unconditional models (e.g., the Unconditional Kindergarten Age in Months Language Model and the Unconditional Kindergarten Instructional Exposure Language Model) to understand which time metric explained a greater proportion of variance in the outcome. Larger *pseudo-r^2* were used to provide preliminary evidence in support of the most effective time metric for each age band and domain of development.

Next, Information Criteria, including model AICs, were used to compare Unconditional Instructional Exposure Growth Models and Unconditional Age in Months Growth Models

(Akaike, 1973 as cited in McCoach et al., 2022). The AIC, which can be calculated by multiplying the number of parameters by two and subtracting 2 times the log-likelihood, provides information about model deviance while imposing a small penalty for less parsimonious models. Smaller AIC values provide evidence of better model fit. Smaller AIC values from each set of non-nested unconditional models were used to provide further evidence of the most effective time metric for each age band and domain of development (McCoach et al., 2022).

Inferring the Nature of the Growth Process

Researchers who have examined longitudinal growth trends over a number of years have frequently found growth rate slows as children age (Mok et al., 2015; Shanley, 2016; Thum & Kuhfeld, 2020). While the proposed study does not seek to draw definitive conclusions about the nature of growth from birth through kindergarten, as assessment records belong to age-separated cohorts of children, the third research question sought to make inferences about the nature or shape of the growth process across domains of learning and development. To achieve this aim, model-estimated linear slopes were examined for children ages birth through kindergarten for each domain of learning and development.

Examining Subgroup Differences in Growth Trajectories

The fourth research question sought to understand how growth trends differed between subgroups of children. For decades, researchers have demonstrated that children's developmental and academic growth trajectories can differ significantly between subgroups of children, such as boys and girls (Hujar et al., 2021; Voyer & Voyer, 2014), White Non-Hispanic, Hispanic, and Black Non-Hispanic children (Kuhfeld et al., 2020; Reardon & Portilla, 2016; Reardon & Galindo, 2009) children who come from low-income or middle- or high-income families (Hujar et al., 2021, Reardon & Portilla, 2016, Roberts & Byrant, 2011), and children that are native and non-native English speakers (Hujar et al., 2021; Roberts & Bryant, 2011). Researchers often examine differences in initial academic abilities and developmental levels (Hujar et al., 2021; Kuhfeld et al., 2020; Reardon & Portilla, 2016) and growth rates over time (Reardon & Galindo, 2009; Voyer & Voyer, 2014) to inform teaching practices and policy decisions (Kuhfeld et al., 2020).

To answer the final research question, child-level characteristics including, gender, race/ethnicity, and primary language were modeled at level two to understand how certain characteristics were related to typical preliminary performance (intercept) and linear growth rate (slope). For this analysis, data were fitted to 72 hierarchical linear growth models, one for each age-band, domain of learning and development, and time metric. Resulting beta coefficients and *p*-values were examined to understand whether child-level characteristics were significantly related to average preliminary performance and mean linear growth rate after controlling for other variables in the model. Beta coefficients with corresponding *p*-values <=.05 were interpreted to understand the typical effect of each child-level characteristic on preliminary developmental status and growth rate over time.

Equation 11 includes The Conditional Level I Instructional Exposure Growth Model, and was used to model the domain-level scaled score for child *i* at time *t*, Y_{ti}, as a function of the child's academic ability or developmental level prior to instruction, π_{0i}, a linear rate of growth, π_{1i}, additional months of instruction since the beginning of the academic year, (*instruction – instruction*$_{1}$)$_{ti}$,, and an error term, e_{ti}. Th$_{e}$ Conditional Level II Instructional Exposure Model includes equations 12 and 13. Equation 12 was used to model each child's developmental level prior to instruction, π_{0i}, as a function of average ability prior to instruction, β_{00}, a female effect, β_{01}, a Spanish effect, β_{02}, an *other* language effect, β_{03}, a Black Non-Hispanic effect, β_{04}, an

Asian Non-Hispanic effect, β_{05}, a Native American, Pacific Islander, Hawaiian Native, and Alaskan Native Non-Hispanic effect, β_{06}, a Bi- and Multiracial effect, β_{07}, a Hispanic effect, β_{08}, and a residual term, r_{0i}. Equation 11 models each child's linear growth rate, π_{1i}, as a function of the average linear growth rate, β_{10}, a female effect, β_{11}, a Spanish effect, β_{12}, an *other* Language effect, β_{13}, a Black Non-Hispanic effect, β_{14}, an Asian Non-Hispanic effect, β_{15}, a Native American, Pacific Islander, Hawaiian Native, and Alaskan Native Non-Hispanic effect, β_{16}, a Bi- and Multiracial effect, β_{17}, a Hispanic effect, β_{18}.

Conditional Level I Instructional Exposure Growth Model

$$Y_{ti} = \pi_{0i} + \pi_{1i}(instruction - instruction_1)_{ti} + e_{ti} \tag{11}$$

Where:

Y_{ti} = academic or developmental level for child *i* at time *t*

π_{0i} = academic or developmental level prior instruction for child *i*

π_{1i} = linear rate of growth for child *i*

instruction = months of instruction for child *i* at time *t*

$instruction_1$ = 0 months of instruction for child *i*

e_{ti} = within child residual at time *t*

Conditional Level II Instructional Exposure Growth Model

(12)

$$\pi_{0i} = \beta_{00} + \beta_{01}(female)_i + \beta_{02}(Spanish)_i + \beta_{03}(other\ language)_i + \beta_{04}(Black) +$$
$$\beta_{05}(Asian)_i + \ \beta_{06}(Native\ American)_i + \ \beta_{07}(Multiple\ Races)_i + \beta_{08}(Hispanic)_i + r_{0i}$$

(13)

$$\pi_{1i} = \beta_{10} + \beta_{11}(female)_i + \beta_{12}(Spanish)_i + \beta_{13}(other\ language)_i + \beta_{14}(Black) + \beta_{15}(Asian)_i + \ \beta_{16}(Native\ American)_i + \ \beta_{17}(Multiple\ Races)_i + \beta_{18}(Hispanic)_i$$

Where:

π_{0i}= academic or developmental level prior to instruction for child *i*

β_{00} = average ability or developmental level prior to instruction

β_{01}= unique female effect in intercept

β_{02}= unique Spanish effect in intercept

β_{03}= unique other language effect in intercept

β_{04}= unique Black Non-Hispanic effect in intercept

β_{05}= unique Asian Non-Hispanic effect in intercept

β_{06}= unique Native American Non-Hispanic effect in intercept

β_{07}= unique Multiracial Non-Hispanic effect in intercept

β_{08}= unique Hispanic effect in intercept

r_{0i}= random error in intercept for child *i*

π_{1i}= linear growth rate for child *i*

β_{10}= average linear growth rate

β_{11}= unique female effect in slope

β_{12}= unique Spanish effect in slope

β_{13}= unique other language effect in slope

β_{14}= unique Black Non-Hispanic effect in slope

β_{15}= unique Asian Non-Hispanic effect in slope

β_{16}= unique Native American Non-Hispanic effect in slope

β_{17}= unique Multiracial Non-Hispanic effect in slope

β_{18}= unique Hispanic effect in slope

Equation 14 includes The Conditional Level I Age in Months Growth Model, and was used to model the domain-level scaled score for child *i* at time *t*, Y_{ti}, as a function of the child's academic ability or developmental level at the beginning of the age band, π_{2i}, a linear rate of growth, π_{3i}, additional months of age from the first age in the age band, $(age - age_1)_{ti}$, and an error term, e_{1ti}. The Conditional Level II Age in Months Model includes equations 15 and 16. Equation 15 was used to model each child's developmental level at the beginning of the age band, π_{2i}, as a function of average ability, β_{20}, a female effect, β_{21}, a Spanish effect, β_{22}, an *other* language effect, β_{23}, a Black Non-Hispanic effect, β_{24}, an Asian Non-Hispanic effect, β_{25}, a Native American, Pacific Islander, Hawaiian Native, and Alaskan Native Non-Hispanic effect, β_{26}, a Bi- and Multiracial effect, β_{27}, a Hispanic effect, β_{28}, and a residual term, r_{0i}. Equation 16 was used to model each child's linear rate of growth, π_{3i}, as a function of the average rate of growth, β_{30}, a female effect, β_{31}, a Spanish effect, β_{32}, an *other* language effect, β_{33}, a Black Non-Hispanic effect, β_{34}, an Asian Non-Hispanic effect, β_{35}, a Native American, Pacific Islander, Hawaiian Native, and Alaskan Native Non-Hispanic effect, β_{36}, a Bi- and Multiracial effect, β_{37}, and a Hispanic effect, β_{38}.

Conditional Level I Age in Months Growth Model

$$Y_{ti} = \pi_{2i} + \pi_{3i}(age - age_1)_{ti} + e_{1ti} \tag{14}$$

Where:

Y_{ti} = academic or developmental level for child *i* at time *t*

π_{2i} = academic or developmental level for child *i* at the youngest age in the age band

π_{3i} = linear rate of growth for child *i*

age = age in months for child *i* at time *t*

age_1 = first age in age band, measured in months

e_{1ti} = within child residual at time *t*

Conditional Level II Age in Months Growth Model

(15)

$$\pi_{2i} = \beta_{20} + \beta_{21}(female)_i + \beta_{22}(Spanish)_i + \beta_{23}(other\ language)_i + \beta_{24}(Black) + \beta_{25}(Asian)_i + \ \beta_{26}(Native\ American)_i + \ \beta_{27}(Multiple\ Races)_i + \beta_{28}(Hispanic)_i + r_{1i}$$

(16)

$$\pi_{3i} = \beta_{30} + \beta_{31}(female)_i + \beta_{32}(Spanish)_i + \beta_{33}(other\ language)_i + \beta_{34}(Black) + \beta_{35}(Asian)_i + \ \beta_{36}(Native\ American)_i + \ \beta_{37}(Multiple\ Races)_i + \beta_{38}(Hispanic)_i$$

Where:

π_{2i}= academic or developmental level for child *i* at the first age in the age band

β_{20} = average ability or developmental level at the first age in the age band

β_{21}= unique female effect in intercept

β_{22}= unique Spanish effect in intercept

β_{23}= unique *other* language effect in intercept

β_{24}= unique Black Non-Hispanic effect in intercept

β_{25}= unique Asian Non-Hispanic effect in intercept

β_{26}= Native American Non-Hispanic effect in intercept

β_{27}= Multiracial Non-Hispanic effect in intercept

β_{28}= Hispanic effect in intercept

r_{1i}= random error in intercept for child *i*

π_{3i}= linear growth rate for child *i*

β_{30}= average linear growth rate

β_{31}= unique female effect in slope

β_{32}= unique Spanish effect in slope

β_{33}= unique *other* language effect in slope

β_{34}= unique Black Non-Hispanic effect in slope

β_{35}= unique Asian Non-Hispanic effect in slope

β_{36}= Native American Non-Hispanic effect in slope

β_{37}= Multiracial Non-Hispanic effect in slope

β_{38}= Hispanic effect in slope

Limitations

Within the context of the current study, there are five data-imposed limitations including, a) ability and growth estimates could not be established for first-, second-, and third-grade children, b) longitudinal data including no more than three assessments per child placed limitations on the structure of growth models and reliability of model-estimated growth parameters, c) assessment data were from age-separated cohorts, which placed limitations on inferences about the nature of the growth process from birth through kindergarten, d) *GOLD*® data tend to include a greater proportion of low-income children than the general population, therefore limiting the generalizability of normative estimates to the broader population of children ages birth through kindergarten, and e) assessment data for Asian children in birth- to one-year-old and kindergarten classrooms was limited, therefore the sample for the present study slightly underrepresents these populations of children.

To begin, ability and growth estimates could not be calculated for first-, second-, and third-grade children even though *GOLD*® was designed and has been externally validated for use with these populations (Lambert 2020; Lambert 2017). The 2021-2022 dataset included less than 2,000 first-, second-, and third-grade children, therefore there were not enough assessment records to establish time-continuous normative scores. If more first-, second-, and third-grade children are assessed using *GOLD*® in subsequent academic years, researchers should seek to replicate research methods used in the present study to derive age-based and instructional-based normative estimates for older children.

Next, longitudinal data including no more than three assessments per child placed limitations on the structure of growth models. Given three measurements per child, only two level-one parameters could be estimated, including, intercept and slope. While more complex MLMs may have demonstrated a better fit to the data, more complex models could not be tested given the current dataset. In the future, if a multi-year longitudinal database is available, researchers may consider fitting longitudinal *GOLD*® data to more complex models such as HLMs with time-varying covariates or MLMs with quadratic parameters to account for acceleration or deceleration in growth rate over time.

Additionally, given only three measurements per child, the reliability of model-estimated parameters was impacted. However, according to *The Standards for Educational and Psychological Measurement*:

> The reliability/precision of measurement is always important. However, the need for precision increases as the consequences of decisions and interpretations grow

> in importance. If a test score leads to a decision that is not easily reversed… a higher degree of reliability/precision is warranted. If a decision can and will be corroborated by information from other sources… scores with more modest reliability/precision may suffice (2014, p. 33).

Given that *GOLD*® data is typically used for a) formative purposes and b) to provide child outcome data as one part of a larger system of accountability (Burts et al., 2016), more modest reliability estimates are likely sufficient. For example, model-estimated intercepts for Unconditional Age in Months Growth Models were reasonably reliable for babies in birth- to one-year-old classrooms (.71-.79) and toddlers in one- to two-year-old classrooms (.76-.81). And model-estimated intercepts for Unconditional Age in Months Growth Models were very reliable for children in two- to three-year-old classrooms (.81-.88), preschool classrooms (.86-.90), pre-k classrooms (.87-.91), and kindergarten classrooms (.90-.94). However, in the presence of more measurements researchers should replicate the current study to increase the reliability/precision of model-estimated parameters and subsequent normative estimates.

Next, children included in the current study belong to age-separated cohorts. While this study sought to examine linear slopes across age-bands and domains of development to infer the nature of the growth process from birth through kindergarten, results should be interpreted with caution as alternate explanations for observable differences between age-separated cohorts of children cannot be ruled out (Singer & Willet, 2013). In the presence of longitudinal data for the same group of children from birth through kindergarten, researchers may reexamine the nature of the growth process using a multilevel modeling approach.

Additionally, although free- and reduced-priced lunch eligibility data was not modeled in the current study because many programs did not enter children's free- and reduced-priced lunch eligibility data, *GOLD*® assessment records tend to overrepresent children from low-income families. Therefore, while results from the present study can likely be generalized to the population of children who are assessed using *GOLD*®, results may not generalize to the broader population of children in the U.S. who are ages birth through kindergarten.

And finally, although the present study sought to use stratified random sampling to obtain a nationally representative sample of children ages birth through kindergarten, there were not enough Asian children to obtain representative samples for two age bands, including birth- to one-year-old children and kindergarten children. While that goal was to obtain samples for each age band that included 162 (5.40%) Asian children, the birth- to one-year-old sample included 134 (4.47%) Asian children and the kindergarten sample included 140 (4.67%) Asian children.

Ethical Considerations

Prior to conducting analyses, the Office of Research Protections and Integrity at the University of North Carolina at Charlotte determined that the current study falls into the exempt category under 45 CFR 46. 104(d) and therefore did not require a full review. Data were obtained legally through an agreement with Teaching Strategies®. There are no foreseen risks to children whose assessment records will be used to derive national normative ability and growth estimates. Results will be shared with Teaching Strategies.

CHAPTER FOUR: RESULTS

The primary purpose of this study was to establish time-continuous normative scores for children ages birth through kindergarten for each major domain of learning and development presented in the *GOLD*® assessment system. Secondary purposes included establishing evidence of the most effective time metric for modeling developmental growth, investigating the nature of developmental growth over time, and examining the relationship between various child-level characteristics and normative scores. Hierarchical linear models were used to investigate research questions. Results corresponding to each research question are provided in subsequent paragraphs.

Hierarchical Linear Model-Estimated Normative Scores

The first research question was: *what do model-based estimates suggest the average child will do at each point in time from birth through kindergarten, across each of the six major domains of learning and development presented in the GOLD® assessment system?* To answer this research question, individual hierarchical linear growth models were used to model each child's domain-level scaled score as a function of age in months and months of instructional exposure. Resulting beta coefficients, including intercepts and linear slopes, were used to construct time-continuous normative estimates for each domain of development, age band, and time metric.

Instructional Exposure Norms

Individual hierarchical linear growth models were used to estimate the average developmental level prior to instruction, β_{00}, and the average linear growth rate, β_{10}, for each age band and domain of development. Next, intercept, β_{00}, and slope, β_{10}, coefficients were used to estimate the typical developmental level for each domain, age band, and month of instruction. Normative scores based on months of instruction are provided for children in two- to three-year-old, preschool, pre-k, and kindergarten classrooms in Table 1. Normative scores based on months of instruction were not created for children in birth- to one-year-old and one- to two-year-old classrooms because the academic year was less predictable for very young children. For example, babies in birth- to one-year-old classrooms frequently enter programs when parental leave ends.

Table 1
Instructional GOLD® Normative Scores

		Average Scaled Score					
Age Band / Grade Level	Months of Instruction	Social Emotional	Physical	Language	Cognitive	Literacy	Math
Two - Three	0	318	419	324	302	354	202
	1	324	427	333	310	360	210
	2	331	435	341	317	367	218
	3	337	443	349	325	374	226
	4	344	450	357	332	380	234
	5	350	458	365	339	387	242
	6	357	466	373	347	394	250
	7	363	474	382	354	401	258
	8	370	482	390	362	407	266
	9	376	490	398	369	414	274
	10	383	498	406	376	421	282
Preschool	0	363	476	381	356	412	273
	1	372	486	392	366	420	283
	2	381	497	402	376	428	292
	3	390	507	413	386	436	301
	4	399	518	423	397	444	310
	5	408	528	434	407	452	319
	6	417	539	444	417	460	328
	7	427	549	455	427	468	337
	8	436	560	465	437	476	347
	9	445	570	475	447	484	356
	10	454	580	486	458	492	365
Pre-K	0	403	521	432	400	457	321
	1	414	534	445	413	465	331
	2	425	547	459	425	474	342

Table 1 (continued)
Instructional GOLD® Normative Scores

		Average Scaled Score					
Age Band / Grade Level	Months of Instruction	Social Emotional	Physical	Language	Cognitive	Literacy	Math
Pre-K	3	437	560	472	438	482	352
	4	448	573	486	451	491	363
	5	460	586	499	464	500	374
	6	471	599	513	477	508	384
	7	482	612	526	489	517	395
	8	494	625	540	502	526	405
	9	505	638	553	515	534	416
	10	516	651	567	528	543	426
Kindergarten	0	457	573	501	464	506	372
	1	469	586	515	478	517	385
	2	482	600	528	492	528	399
	3	495	613	542	506	539	412
	4	507	627	555	520	550	426
	5	520	640	569	535	561	440
	6	533	654	583	549	573	453
	7	545	667	596	563	584	467
	8	558	681	610	577	595	480
	9	571	694	623	592	606	494
	10	583	708	637	606	617	507

Normative estimates provided in Table 1 can be used to understand the typical developmental level, as measured by *GOLD*®*,* for each month of instruction for any child who is in a two- to three-year-old, preschool, pre-k, or kindergarten program. Instructional norms can be used to understand relative performance, given age band and months of instruction. For example, a pre-k teacher who finalized assessment scores in mid-October may review the row that corresponds to pre-k and three months of instruction. The normative estimates for this row are social-emotional (437), physical (560), language (472), cognitive (438), literacy (482), and mathematics (352). Next, she can compare her children's domain-level scaled scores to the national normative estimates to understand relative performance. By engaging in this process, she can begin to make decisions about children in her classroom who may need additional monitoring, support or extensions.

Normative estimates outlined in Table 1 can also be used to understand typical growth between any two months of instruction within a given age band. For example, a kindergarten teacher who finalizes fall ratings after one month of instruction, and winter ratings after five months of instruction can see that kindergarteners typically gain 44 literacy scale points, 51 social-emotional scale points, 54 physical, language, and mathematics scale points, and 56 cognitive scale points between measurements. She can use this information to understand if children in her classroom are developing at similar rates as kindergarteners from a nationally representative sample of children. By engaging in this process, she can make decisions about children who may need additional monitoring or support to make adequate gains over time.

Validity of Estimates. According to *The Standards for Educational and Psychological Measurement*, "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (2014, p. 11). Within the context of the present study, the term validity is used to describe the extent to which evidence suggests that model-parameter estimates, and subsequent national normative scores are accurate and stable for children ages birth through kindergarten. While valid and reliable normative estimates alone do not fully substantiate the norm-referenced validity argument, valid and reliable normative scores are prerequisite to supporting teachers and administrators in making useful, meaningful, and appropriate inferences about children's relative performance. Therefore, the validity and reliability of model-based estimates and corresponding normative scores were investigated in several ways. First, the reliability of model-estimated intercepts was examined. Results suggested that model-estimated instructional exposure intercept values were reliable across domains of development for children in two- to three-year-old classrooms (.81-.89) and very reliable for children in preschool classrooms (.86-.90), pre-k classrooms (.85-.89), and kindergarten classrooms (.87-.93). See Table 2 for reliability estimates for Unconditional Instructional Exposure Model-estimated intercepts for children in two- to three-year-old, preschool, pre-k, and kindergarten classrooms.

Table 2
Reliability of Instructional Exposure Intercepts

Domain	Two to Three	Preschool	Pre-K	Kindergarten
Social Emotional	0.83	0.87	0.86	0.87
Physical	0.81	0.86	0.85	0.88
Language	0.89	0.90	0.89	0.91
Cognitive	0.84	0.86	0.88	0.90
Literacy	0.83	0.86	0.85	0.93
Mathematics	0.85	0.88	0.89	0.90

Next, we would expect children's scores to increase across grade levels and months of instruction. This pattern of growth is generally observable across estimates provided in Table 1. For example, if we look at the progression of language development from the time a child is in a two- to three-year-old classroom through kindergarten, the typical child in a two- to three-year-old classroom begins the year with a scaled score of 324 (zero months of instruction) and ends the year with a scaled score of 398-406 (9-10 months of instruction). A typical preschooler begins the year with a scaled score of 381 and ends the year with a scaled score of 475-486. A typical child in pre-k begins the year with a scaled score of 432 and ends the year with a scaled score of 553-567. And a typical kindergartener begins the year with a scaled score of 501 and ends the year with a scaled score of 623-637. While there is some overlap between contiguous grade-levels, as children progress through age bands and months of instruction, scores generally increase.

Finally, model-based estimates were compared to the normative scores provided in the *Technical Manual for the Teaching Strategies GOLD® Assessment (2nd Edition).* While the norms provided in the manual present challenges for teachers who finalize assessments earlier or later than average, they still generally reflect typical developmental status at the average fall, winter, and spring assessment finalization dates. HLM-based estimates for the most typical months for finalizing scores were compared to the estimates derived using score distributions. Results suggested HLM-based estimates and score distribution-based estimates provided very similar and at times identical results. For example, in 2021, assessment scores for kindergarteners were most frequently finalized after two months of instruction, six months of instruction, and nine months of instruction. HLM-based social-emotional normative estimates were 482 (two months of instruction), 533 (six months of instruction), and 571 (nine months of instruction). Similarly, score-distribution based social-emotion normative scores for kindergarteners were 483 (fall), 527 (winter), and 569 (spring). In this example, estimates displayed discrepancies of 1, 6, and 2 scale points for fall, winter, and spring measurement occasions, respectively. On average, the discrepancy in points between traditional estimates and HLM-based estimates across domains and assessment occasions was 6.44 points for two- to three-year-old children, 6.22 points for preschoolers, 5.67 points for pre-k children, and 4.83 points for kindergarten children. Together, these results suggest that HLM-based estimates provide reasonably valid normative estimates. See Table 3 for a full comparison of normative scores derived using score distributions and HLMs.

Table 3

Comparison of Score Distribution and HLM-Based Instructional Norms

		50th Percentile Scaled Score											
		Social Emotional		Physical		Language		Cognitive		Literacy		Mathematics	
Age Band / Grade Level	Trimester	HLM	Score Dist.	HLM	Score Dist.	HLM	Score Dist.	HLM	Score Dist.	HLM	Score Dist.	HLM	Score Dist.
Two - Three	Fall	331	331	435	428	341	348	317	319	367	377	218	229
	Winter	354	354	462	470	369	372	343	351	391	398	246	257
	Spring	373	376	485	494	394	395	366	369	411	425	270	280
Preschool	Fall	381	381	497	506	402	410	376	381	428	443	292	298
	Winter	413	420	534	543	439	445	412	415	456	470	324	331
	Spring	441	442	565	568	470	473	442	448	480	487	352	354
Pre-K	Fall	425	431	547	556	459	473	425	442	474	485	342	350
	Winter	466	467	593	593	506	515	471	475	504	509	379	384
	Spring	500	499	632	635	547	551	509	509	530	531	412	408
Kindergarten	Fall	482	483	600	615	528	524	492	492	528	529	399	392
	Winter	533	527	654	668	583	585	549	545	573	573	453	446
	Spring	571	569	694	710	623	627	592	592	606	606	494	490

Note. The most typical months for finalizing assessment scores for children in two- to three-year-old classrooms, preschool, and pre-k classrooms were after 2 months (October), 5.5 months (January/February), and 8.5 months (April/May) of instruction. The most typical months for finalizing assessment scores for children in kindergarten classrooms were after 2 months (October), 6 months (February), and 9 months (May).

Age Norms

Individual hierarchical linear growth models were used to estimate the average developmental level for the youngest children in each age band, β_{20}, and the average linear growth rate, β_{30}, for each age band and domain of development. Next, intercept, β_{20}, and slope, β_{30}, coefficients were used to estimate the typical developmental level for each domain, age band, and month of age. Norms are provided for every three months of age in Table 4; however, intercept and slope coefficients could be used to derive a normative ability estimate for any month of age.

Table 4
Age GOLD® Normative Scores

		Average Scaled Score					
Age Band/Grade Level	Age in Months	Social Emotional	Physical	Language	Cognitive	Literacy	Math
Birth - One Year	0	104	76	67	58	0	0
	3	138	131	103	95	40	0
	6	173	187	139	131	95	12
	9	207	242	174	168	149	32
	12	241	297	210	205	204	53
	15	275	352	246	241	259	73
One - Two Years	12	240	311	202	203	232	57
	15	261	338	231	228	261	91
	18	282	364	261	253	291	124
	21	303	391	291	277	320	158
	24	323	418	320	302	349	191
	27	344	445	350	327	378	225
Two - Three Years	24	299	397	300	281	334	179
	27	318	419	324	303	354	202
	30	337	442	348	324	373	225
	33	355	464	371	345	392	248
	36	374	487	395	366	412	271
	39	392	509	419	387	431	294
Preschool	36	327	435	337	316	381	236
	39	351	463	366	343	402	261
	42	376	491	395	371	424	286
	45	400	519	424	398	445	311
	48	425	547	453	425	467	336
	51	449	575	482	452	488	361
	54	474	603	511	480	510	386
Pre-K	48	363	476	381	353	427	280
	51	393	510	417	387	449	309
	54	423	544	454	421	472	338
	57	452	578	490	456	494	367
	60	482	612	526	490	517	395
	63	512	645	563	524	540	424
	66	541	679	599	558	562	453

Table 4 (continued)
Age GOLD® Normative Scores

Age Band/Grade Level	Age in Months	Average Scaled Score					
		Social Emotional	Physical	Language	Cognitive	Literacy	Math
Kindergarten	60	413	527	450	411	462	321
	63	447	563	488	450	494	359
	66	481	599	525	489	525	396
	69	515	635	563	528	556	433
	72	549	671	600	567	587	471
	75	582	707	638	606	618	508
	78	616	742	676	645	650	545

Normative estimates provided in Table 4 can be used to understand the typical developmental level, as measured by *GOLD®*, for any reasonably aged child in a birth- to one-year-old, one- to two-year-old, two- to three-year-old, preschool, pre-k, or kindergarten classroom that uses *GOLD®*. Furthermore, norms provided in Table 4 can be used to understand relative performance, given a child's age band or grade level and age in months. Normative scores based on age, rather than months of instruction, may be particularly useful for teachers with children who are younger or older than their peers. For example, a kindergarten teacher may have children as young as 60 months and as old as 69 months at the beginning of the academic year. After finalizing the fall assessment scores, she may wonder if the discrepancies in language scaled scores between her youngest and oldest children are concerning or reflective of typical patterns of development. To answer her question, she can locate the norms for kindergarten language development that correspond to 60 months of age (450 points) and 69 months of age (515 points).

Normative scores based on age may also be particularly useful for teachers in birth- to one-year-old and one- to two-year-old classrooms where development occurs rapidly and is closely related to biological age (Harkness et al., 2013). For example, a teacher in a birth- to one-year-old classroom could use the age-based normative scores to understand whether the babies in her classroom are developing as expected. For instance, she can compare the three-month-old babies' domain-level scaled scores to the national normative scores that correspond to children in birth- to one-year-old classrooms and three months of age. The normative estimates for this row are social-emotional (138), physical (131), language (103), cognitive (95), literacy (40), and mathematics (0). Next, she can compare her babies' domain-level scales scores to the national normative estimates to understand relative performance. By engaging in this process, she can make decisions about which babies may need additional support to meet developmental milestones and effectively communicate with families and other stakeholders.

Finally, normative estimates provided in Table 4 can also be used to understand typical growth between any two months of age within the same age band. For example, a teacher in a birth- to one-year-old classroom who acquires a new 12-month-old toddler may be interested to know how much physical growth the child might make before leaving her classroom around the time she's 15 months. She can use the norms provided in Table 4 to see that the child may gain about 55 physical scale points before leaving her classroom.

Validity of Estimates. Once again, while valid and reliable normative estimates alone do not fully substantiate the norm-referenced validity argument, valid and reliable normative scores are prerequisite to supporting teachers and administrators in making useful, meaningful, and appropriate inferences about children's relative performance. Therefore, the validity and reliability of model-based estimates and corresponding normative scores were investigated in several ways.

First, the reliability of model-estimated intercepts was examined. Results suggested that model-estimated age intercept values were reliable for children in birth- to one-year-old classrooms (.71-.81), one- to two-year-old classrooms (.76-.81), and two- to three-year-old classrooms (.82-.88) and very reliable for children in preschool (.86-.90), pre-k (.87-.91), and kindergarten (.90-.94) classrooms. See Table 5 for reliability estimates for Unconditional Age in Months Model-estimated intercepts for children in birth- to one-year-old, one- to two-year-old, two- to three-year-old, preschool, pre-k, and kindergarten classrooms.

Table 5
Reliability of Age in Months Intercepts

Domain	Birth - One	One - Two	Two - Three	Preschool	Pre-K	Kinder
Social-Emotional	0.78	0.78	0.83	0.87	0.88	0.90
Physical	0.77	0.76	0.81	0.87	0.87	0.90
Language	0.81	0.81	0.88	0.90	0.91	0.93
Cognitive	0.79	0.78	0.84	0.87	0.90	0.92
Literacy	0.71	0.76	0.82	0.86	0.87	0.94
Mathematics	0.75	0.77	0.85	0.88	0.91	0.93

Next, we would expect scaled scores to increase as children get older. This pattern is generally observable for each domain of development as we move down the columns. For example, the typical social-emotional scaled scores for the youngest children in each age band are: 104 (birth- to one-year-olds), 240 (one- to two-year-olds), 299 (two- to three-year-olds), 327 (preschool), 363 (pre-k), and 413 (kindergarten). See Table 4 to review the progression of normative scores as children age for every domain of learning and development.

Finally, we would expect that Unconditional Age in Months Growth Model-estimated intercepts would be lower than Unconditional Instructional Exposure Model-estimated intercepts because age in month intercept values are generally reflective of typical development for the youngest child prior to instruction and instructional exposure intercept values are generally reflective of the typical development for the average aged child prior to instruction. This pattern can be observed for every domain of development and age band. For example, the expected mathematics scaled score for a kindergartener who is 60 months old is 321, while the expected scaled score for a kindergartener prior to instruction is 372.

The Most Effective Time Metric

The second research question was: *which time metric is most effective for modeling growth, approximate instructional exposure or age in months, from birth through kindergarten?* To answer this research question, two strategies were used. First, *pseudo-r^2* statistics were calculated for each unconditional model to understand which time metric explained a greater proportion of the level-one variance when compared to the completely unconditional model. Next, AIC values were calculated for Unconditional Instructional Exposure Growth models and Unconditional Age in Months Growth models. For every age band and domain, AIC values for pairs of non-nested models were compared to determine which model demonstrated better fit to the data. Although resulting *pseudo-r^2* statistics and AIC values were very similar across every set of non-nested models due to highly correlated ($r = .68$) and linearly measured time variables, if fit statistics agreed, preliminary inferences were made about the most effective time metric.

In the past researchers have found that babies' developmental growth is closely related to biological age (Harkness et al., 2013), yet other researchers have documented how children as young as two- or three-years old are significantly influenced by the environment (Yoshikawa et al., 2013) and interactions with caregivers (Chetty et al., 2011; Gialamas et al., 2013). Given this research, it was hypothesized that age in months may be the most effective predictor of developmental growth for children in the youngest two age bands across domains of development and instructional exposure may be the most effective predictor for children in two- to three-year-old classrooms and beyond.

Birth to One Year

Despite the fact that researchers have identified a strong and positive relationship between babies' biological age and developmental growth (Harkness et al., 2013; Rebello-Britto et al., 2013), the most effective predictor of developmental growth for babies in birth- to one-year-old classrooms was inconsistent across domains of learning and development. Together, larger *pseudo-*r^2 statistics and smaller AIC values suggested that age in months was the most effective time metric for modeling babies' social-emotional and physical development. Conversely, larger *pseudo-*r^2 statistics and smaller AIC values indicated that instructional exposure was the most effective predictor of babies' mathematics development. And finally, given equivalent or larger *pseudo-*r^2 statistics for Unconditional Language, Cognitive, and Literacy Age in Months Growth Models and smaller AIC values for Unconditional Language, Cognitive, and Literacy Models, no inferences were made about the most effective time metric for these domains of learning and development. Results, including level-one *pseudo-*r^2 and AIC values are provided in Table 6.

Table 6

Comparison of Unconditional Model Fit Statistics for the First Three Age Bands

		Birth to One		One to Two		Two to Three	
Domain	Model Fit	Instruct.	Age	Instruct.	Age	Instruct.	Age
Social-Emotional	*pseudo-r^2*	0.365	0.366	0.190	0.191	0.134	0.135
	AIC	93295	92872	92938	92935	93452	93534
Physical	*pseudo-r^2*	0.447	0.448	0.175	0.176	0.116	0.117
	AIC	94870	94005	93947	93927	94425	94490
Language	*pseudo-r^2*	0.357	0.355	0.238	0.239	0.119	0.120
	AIC	90952	90586	92474	92361	94002	94018
Cognitive	*pseudo-r^2*	0.375	0.375	0.230	0.231	0.152	0.153
	AIC	93992	93417	93349	93203	94435	94443
Literacy	*pseudo-r^2*	0.278	0.276	0.157	0.159	0.094	0.095
	AIC	100741	100468	94913	94888	92255	92303
Mathematics	*pseudo-r^2*	0.128	0.127	0.231	0.232	0.143	0.143
	AIC	93632	93723	95538	95338	93779	93791

Note. Green indicates age in months was the most effective time metric as evidenced by larger *pseudo-r^2 statistics* and smaller AICs, blue indicates that instructional exposure was the most effective time metric as evidenced by larger *pseudo-r^2 statistics* and smaller AICs, and red indicates that model fit statistcs were contradictory.

One to Two Years

Although fit statistics were similar across pairs of non-nested models, results, including larger *pseudo-r^2* statistics and smaller AIC values, indicated that age in months was the most effective time metric for modeling developmental and academic growth for toddlers in one- to two-year-old classrooms. For every domain of learning and development, the proportion of within-child variance was reduced more significantly by modeling domain-level scaled scores as a function of age, rather than instruction. For example, by modeling language development as a function of age rather than instruction, the level-one residual variance was reduced by 23.9% (*pseudo-r^2*$_{\text{age}}$ = .239) rather than 23.8% (*pseudo-r^2*$_{\text{instruction}}$ = .238). Furthermore, smaller AIC values also suggested that age in months was the most effective predictor of developmental growth for this age group, regardless of domain. For example, the AIC was smaller for the Unconditional Language Age in Months Growth Model (AIC $_{\text{age}}$ = 92,361) than the Unconditional Language Instructional Exposure Growth Model (AIC $_{\text{instruction}}$ = 92,474). Additional level-one *pseudo-*r^2 statistics and AIC values for one- to two-year-old models are provided in Table 6.

Two to Three Years

Researchers have found that caregiver interactions (Gialamas et al., 2013), teacher quality (Chetty et al., 2011), and the environment (Rebello-Britto et al., 2013) profoundly influence children's development beginning around two or three years of age. Given these findings, it was hypothesized that the most effective time metric for modeling children's growth may change from age to instructional exposure for children in two- to three-year-old classrooms. However, results, including model fit statistics for two- and three-year-old children, presented a less clear pattern. Across domains of development, larger *pseudo-*r^2 statistics suggested that age was the more effective predictor of developmental growth. Yet, smaller AIC values suggested that instructional exposure was the more effective predictor of growth. While definitive conclusions cannot be made about conflicting results, possible explanations include a) age and instruction are equal predictors of developmental growth for children in two- to three-year-old classrooms or b) results are a product of linearly measured and strongly correlated time variables.

Preschool, Pre-k, and Kindergarten

As children get older, research suggests that relationships with caregivers (Gialamas et al., 2013), teacher quality (Chetty et al., 2011), and environmental factors (Rebello-Britto et al., 2013) have a greater influence on developmental growth. Therefore, it was hypothesized that instructional exposure would be a more effective predictor of developmental growth for children in preschool, pre-k, and kindergarten classrooms. Although fit statistics were similar across pairs of non-nested models, results, including larger *pseudo-*r^2 statistics and smaller AIC values, generally indicated that instructional exposure was the most effective time metric for modeling developmental and academic growth for children in preschool, pre-k, and kindergarten classrooms. For almost every domain of learning and development, the proportion of within-child variance was reduced more significantly by modeling domain-level scaled scores as a function of instructional exposure, rather than age. For example, when preschoolers' literacy development was modeled as a function of instruction rather than age, the level-one residual variance was reduced by 17.6% (*pseudo-*$r^2_{\text{instruction}}$ = .176) rather than 17.5% (*pseudo-*r^2_{age} = .175). Similarly, when pre-k children's literacy development was modeled as a function of instructional exposure rather than age, the level-one residual variance was reduced by 28.8% (*pseudo-*$r^2_{\text{instruction}}$ = .288) as opposed to 28.4% (*pseudo-*r^2_{age} = .284). And finally, when kindergarteners' literacy development was modeled as a function of instruction rather than age, the level-one residual variance was reduced by 43.1% (*pseudo-*$r^2_{\text{instruction}}$ = .431) rather than 42.8% (*pseudo-*r^2_{age} = .428).

Lower AIC values for Unconditional Instructional Exposure Growth Models also suggested that instructional exposure was a better predictor of developmental growth for every domain of learning and development for preschool, pre-k, and kindergarten children. For example, AIC values for Unconditional Literacy Instructional Exposure Growth Models were lower ($\text{AIC}_{\text{preschool}}$ = 90,069; $\text{AIC}_{\text{pre-k}}$ = 88,280; $\text{AIC}_{\text{kindergarten}}$ = 86,464) than Unconditional Literacy Age in Months Growth Models ($\text{AIC}_{\text{preschool}}$ = 90,210; $\text{AIC}_{\text{pre-k}}$ = 88,881; $\text{AIC}_{\text{kindergarten}}$ = 87,400). Additional level-one *pseudo-*r^2 statistics and AIC values for preschool, pre-k, and kindergarten models are provided in Table 7.

Table 7

Unconditional Model Fit Statistics for Preschool, Pre-K, and Kindergarten

		Preschool		Pre-K		Kindergarten	
Domain	Model Fit	Instruct.	Age	Instruct.	Age	Instruct.	Age
Social-Emotional	*pseudo-r^2*	0.194	0.192	0.310	0.305	0.415	0.415
	AIC	94736	94928	94208	95012	91440	92459
Physical	*pseudo-r^2*	0.167	0.166	0.288	0.283	0.271	0.270
	AIC	94099	94274	93812	94477	93507	94055
Language	*pseudo-r^2*	0.148	0.147	0.257	0.255	0.315	0.313
	AIC	95880	96007	96028	96577	93994	94600
Cognitive	*pseudo-r^2*	0.205	0.204	0.313	0.309	0.422	0.419
	AIC	95648	95882	94378	95131	93506	94481
Literacy	*pseudo-r^2*	0.176	0.175	0.288	0.284	0.431	0.428
	AIC	90069	90210	88280	88881	86464	87400
Mathematics	*pseudo-r^2*	0.198	0.198	0.327	0.325	0.476	0.472
	AIC	92875	93010	90637	91328	88624	89809

Note. Blue indicates that instructional exposure was the most effective time metric as evidenced by larger *pseudo-r^2 statistics* and smaller AICs.

Patterns Across Age Bands

Although this research question can best be answered by comparing model fit statistics for each set of non-nested unconditional models, model fit statistics revealed two other interesting patterns that may aid in contextualizing results. First, while there were frequently differences in explanatory power between time metrics across unconditional models, differences were modest. For example, instructional exposure generally accounted for 0.2%-0.5% more of the variability in pre-k children's developmental growth than age. Similarly, while Pre-K Unconditional Instructional Exposure Growth Model AICs were consistently lower than Pre-K Unconditional Age Growth Model AICs, differences were small. This pattern of results can be explained by examining the way time variables were measured and modeled. While age in months was obtained by subtracting a child's birth date from each assessment date, months of instructional exposure was obtained by subtracting each child's assessment date from a typical or average school start date, August 14, 2021. Therefore, time metrics were both linearly measured and strongly correlated (r = .68).

Next, the explanatory power of both time metrics varied significantly across age bands. Relatively large level-one *pseudo-*r^2 statistics suggested that both approximate instructional exposure and age in months explained a significant proportion of variation in developmental growth across most domains of development for children in birth- to one-year-old classrooms, pre-k classrooms, and kindergarten classrooms. For example, 47.6% of the variance in kindergarteners' mathematics scores could be explained by instructional exposure (*pseudo-*r^2 = .476) and 47.2% of the variance in kindergartener's mathematic scores could be explained by age (*pseudo-*r^2 = .472). Conversely, relatively small level-one *pseudo-*r^2 statistics suggested that both approximate instructional exposure and age in months did not explain a significant proportion of variation in developmental growth across domains for children in one- to two-year-old, two- to three-year old, and preschool classrooms. For example, only 17.6% of the variation in preschoolers' literacy scores could be explained by instructional exposure (*pseudo-*r^2 = .176) and 17.5% of the variation in preschoolers' literacy scores could be explained by age (*pseudo-*r^2 = .175). While small *pseudo-*r^2 statistics indicate that there is a significant proportion of residual level-one variance, and residual variance by definition is unexplained error, differences in explanatory power between age bands could be due to more severe rater effects at particular grade levels.

Nature of Developmental Growth

The third research question was: *what do hierarchical linear model-estimated slopes from age-separated cohort data suggest about the shape of the developmental pathway from* *birth through kindergarten across domains of learning and development?* To answer this research question, model-estimated linear slopes were examined sequentially across age bands and time metrics. Linear slopes, β_{30}, from Unconditional Age Models were examined successively from birth through kindergarten to make inferences about the nature of developmental growth over time for each domain of development.

Research suggests that children's brains develop more rapidly from ages birth to five, than at any other point in the lifespan (Harkness et al., 2013; Rebello-Britto et al., 2013). Furthermore, researchers have widely observed that developmental and academic growth slows as children age and progress through subsequent grade levels (Lee, 2010; Mok et al., 2016; Shanley, 2016; Shin et al., 2013). Given these widely observed patterns of development, it was hypothesized that linear slopes, regardless of time metric or domain of development, would be the steepest for children in the first age band and demonstrate a pattern of deceleration for subsequent age bands.

Age

While data belong to age-separated cohorts of children and definitive conclusions cannot be made about the nature of growth from birth through kindergarten, model-estimated linear slopes, β_{30} suggested that developmental growth rate decelerated from birth to age three across most domains of development. For example, babies in birth- to one-year-old classrooms gained on average, 11.38 (p <.001) social-emotional scale points per month of age, toddlers in one- to two-year-old classrooms gained on average, 6.93 (p <.001) social-emotional scale points per month of age, and children in two to three-year-old classrooms gained on average, 6.19 (p <.001) social-emotional scale points per month of age. This pattern of deceleration in growth rate from birth to age three was consistent across social-emotional, physical, language, cognitive, and literacy domains. While this initial sequence provided support for the hypothesis that growth rate decelerates as children age, developmental growth rate began to accelerate for preschool, pre-k, and kindergarten children. For example, preschoolers earned on average, 8.18 (*p* <.001) social-emotional scale points per month of age, pre-k children earned on average, 9.90 (*p* <.001) social-emotional scale points per month of age, and kindergarten children earned on average 11.30 (*p* <.001) social-emotional scale points per month of age. This pattern of acceleration was consistent across all domains of learning and development.

Again, although definitive conclusions cannot be drawn about the nature of growth from birth through kindergarten, there are three plausible explanations for the observable pattern. First, differences in growth rates between age-separated cohorts of children could be due to differences in cohort characteristics. Second, more rapid growth rates for the first (birth through one year) and sixth (kindergarten) age bands could be reflective of the curvilinear logistic function used to convert raw scores to scaled scores where smaller raw score gains at the tails of the distribution contribute to larger differences in scaled scores. And third, it's possible that growth rate, as measured by *GOLD*®, slows across the first three age bands due to the natural slowing of development as children age (Lee, 2010; Shanley, 2016) and strengthens across the next three age bands as instruction intensifies. See Table 8 for Unconditional Age in Months Growth Model-estimated linear slopes, β_{30}, and corresponding *p*-values.

Table 8

GOLD® Scaled Score Points Per Month of Age

	Social Emotional		Physical		Language		Cognitive		Literacy		Mathematics	
	β_{30}	p	β_{30}	p	β_{30}	p	β_{30}	p	β_{30}	p	β_{30}	p
Birth to One	11.38	<0.001	18.43	<0.001	11.90	<0.001	12.20	<0.001	18.30	<0.001	6.82	<0.001
One to Two	6.93	<0.001	8.92	<0.001	9.86	<0.001	8.21	<0.001	9.71	<0.001	11.16	<0.001
Two to Three	6.19	<0.001	7.49	<0.001	7.96	<0.001	7.07	<0.001	6.45	<0.001	7.71	<0.001
Preschool	8.18	<0.001	9.36	<0.001	9.63	<0.001	9.08	<0.001	7.16	<0.001	8.34	<0.001
Pre-K	9.90	<0.001	11.26	<0.001	12.16	<0.001	11.44	<0.001	7.54	<0.001	9.58	<0.001
Kindergarten	11.30	<0.001	11.94	<0.001	12.54	<0.001	12.99	<0.001	10.40	<0.001	12.43	<0.001

Instructional Exposure

Similarly, model-estimated linear slopes, β_{10}, suggested a pattern of acceleration for children in two- to three-year-old classrooms through kindergarten across domains of development. For example, in mathematics, children in two- to three-year-old classrooms gained on average, 7.99 (p <.001) mathematics scale points per month, preschoolers gained on average, 9.16 (p <.001) mathematics scale points per month, pre-k children gained on average, 10.59 (p <.001) mathematics scale points per month, and kindergarteners gained on average, 13.56 (p <.001) mathematics scale points per month. While results were not necessarily consistent with the broader literature, which suggests that developmental growth generally slows as children progress through subsequent grade levels (Lee, 2010; Shanley, 2016), results were consistent with the 50th percentile gain scores provided in the *Technical Manual for the Teaching Strategies GOLD® Assessment (2nd Edition).* For example, children in two- to three-year-old classrooms gained on average, 54.85 mathematics scale points from the average fall to spring measurement (8.43 scale points per month of instruction), preschoolers gained on average 62.56 mathematics scale points from the fall to spring measurement (9.62 scale points per month of instruction), children in pre-k gained on average 63.96 mathematics scale points (9.84 scale points per month of instruction), and kindergarteners gained on average 95.32 mathematics scale points (13.62 scale points per month instruction) (Lambert, 2020).

Again, although definitive conclusions cannot be made about the nature of growth from age two through kindergarten, there are three plausible explanations for accelerating growth rates across age bands. First, differences in growth rates between age-separated cohorts of children could be due to differences in cohort characteristics. Second, acceleration in developmental growth rates from the third (two- to three-years-old) through sixth (kindergarten) age bands could be reflective of the curvilinear logistic function used to convert raw scores to scaled scores. And third, it's possible that growth rate, as measured by *GOLD*® strengthens across age bands as instruction intensifies for preschool, pre-k, and kindergarten children. See Table 9 for Unconditional Instructional Exposure Growth Model-estimated linear slopes, β_{10}, and corresponding p-values.

Table 9

GOLD® Scaled Score Points Per Month of Instruction

	Social Emotional		Physical		Language		Cognitive		Literacy		Mathematics	
	β_{10}	p	β_{10}	p	β_{10}	p	β_{10}	p	β_{10}	p	β_{10}	p
Two to Three	6.53	<0.001	7.88	<0.001	8.16	<0.001	7.39	<0.001	6.70	<0.001	7.99	<0.001
Preschool	9.12	<0.001	10.45	<0.001	10.47	<0.001	10.18	<0.001	8.00	<0.001	9.16	<0.001
Pre-K	11.36	<0.001	12.98	<0.001	13.49	<0.001	12.82	<0.001	8.66	<0.001	10.59	<0.001
Kindergarten	12.66	<0.001	13.50	<0.001	13.59	<0.001	14.23	<0.001	11.08	<0.001	13.56	<0.001

Subgroup Differences

The fourth research question was: *How do growth trends differ between subgroups of children (e.g., by gender, primary language, and race/ethnicity)?* To understand how each characteristic affected the expected growth trajectory, including preliminary performance and growth rate, child-level characteristics were dummy coded and modeled at level two. Resulting beta coefficients and p-values were used to understand the magnitude and significance of each effect. Results generally suggested that girls, children whose primary language was not English, Black Non-Hispanic, and Hispanic children experienced significantly different patterns of performance when compared to the reference group (White Non-Hispanic boys who speak English). Conversely results generally suggested that Asian Non-Hispanic children, Native American Non-Hispanic children, and Bi- or Multiracial Non-Hispanic children did not experience significantly different patterns of growth across age bands and domains of development. Results for subgroups of children who demonstrated significantly different patterns of performance and/or growth are discussed in greater detail in subsequent paragraphs.

Girls

Research generally suggests that girls demonstrate stronger abilities across domains of development at school entry (Kuhfeld et al., 2020; Voyer & Voyer, 2014) that persist over time (Voyer & Voyer, 2014). Therefore, it was hypothesized that girls would demonstrate higher levels of preliminary development and stronger growth over time. Results, including positive intercept beta coefficients, β_{01} and β_{21}, and statistically significant *p*-values provided support for study hypotheses and agreed with the broader body of literature on this topic. Generally, girls in two- to three-year-old and beyond demonstrated stronger preliminary developmental levels across domains of development prior to instruction, β_{01}, when compared to boys. For example, girls in two- to three-year-old classrooms earned on average, 18.74 (*p* <.001) additional language scale points prior to instruction. Girls continued to demonstrate stronger language abilities prior to instruction throughout preschool (β_{01} = 23.96, *p* < .001), pre-k (β_{01} = 16.03, *p* < .001) and kindergarten (β_{01} =13.65, *p* < .001). Similarly, girls in two- to three-year-old classrooms who were 24 months of age at the beginning of the academic year earned on average, 17.72 (*p* <.001) additional language scale points, when compared to boys who were the same age. Girls continued to demonstrate stronger levels of language development throughout preschool (β_{21} = 25.77, *p* < .001), pre-k (β_{21} = 16.58, *p* < .001), and kindergarten (β_{21} = 14.01, *p* < .001) when compared to boys who were the same age. Together these results suggest that gaps in development between boys and girls begin around age two and persist across almost every domain of learning and development.

Finally, despite previous studies which have demonstrated that girls tend to exhibit stronger gains over time across domains (Voyer & Voyer, 2014), results from the present study suggested that there were not any systematic patterns of differences in growth rates across age bands, domains of development, and time metrics. However, kindergarten girls demonstrated significantly stronger growth rates, when compared to kindergarten boys. Positive social-emotional, language, cognitive, and literacy slope coefficients, β_{11}, and statistically significant *p*-values suggested that girls made stronger gains for each month of instruction. For example, kindergarten girls gained on average an additional .69 (*p* = .007) mathematics scale points per month, when compared to boys. Similarly, positive literacy and mathematics slope coefficients, β_{31}, and statistically significant p-values suggested that girls made stronger gains for each additional month of age. For example, kindergarten girls gained on average an additional .56 (*p* = .021) mathematics scale points per month of age, when compared to boys. Together, these results suggest that gaps in performance between boys and girls may begin to widen around the time children are five or six years old. See Tables 10 and 11 for female effects on instructional growth parameters and age growth parameters respectively. Positive and statistically significant beta coefficients are highlighted in green to emphasize favorable effects. Conversely, negative and statistically significant beta coefficients are highlighted in red to emphasize adverse effects.

Table 10

Female Effect on Instructional Growth Parameters

	Birth to One		One to Two		Two to Three		Preschool		Pre-K		Kindergarten	
	β	p	β	p	β	p	β	p	β	p	β	p
Intercept, π_0												
Social β_{01}	0.19	0.937	1.14	0.622	11.90	<0.001	14.92	<0.001	13.32	<0.001	12.08	<0.001
Phys. β_{01}	1.48	0.673	-1.56	0.608	11.04	<0.001	7.56	0.030	10.37	0.001	6.21	0.106
Lang. β_{01}	3.38	0.173	5.96	0.025	18.74	<0.001	23.96	<0.001	16.03	<0.001	13.65	<0.001
Cog. β_{01}	2.56	0.316	-1.23	0.597	11.75	<0.001	13.23	<0.001	11.54	<0.001	10.37	<0.001
Lit. β_{01}	7.43	0.110	3.30	0.389	14.36	<0.001	17.42	<0.001	9.69	<0.001	5.64	0.002
Math β_{01}	5.29	0.018	7.28	0.025	12.55	<0.001	14.41	<0.001	8.40	<0.001	1.97	0.386
Slope, π_1												
Social β_{11}	0.39	0.193	0.425	0.158	-0.11	0.711	0.40	0.227	0.05	0.876	0.68	0.015
Phys. β_{11}	0.21	0.614	0.54	0.172	-0.63	0.122	0.59	0.167	-0.09	0.816	-0.06	0.876
Lang. β_{11}	0.34	0.264	0.37	0.266	-0.56	0.112	-0.11	0.783	0.03	0.936	0.50	0.116
Cog. β_{11}	0.16	0.586	0.77	0.012	-0.41	0.185	0.39	0.289	0.41	0.225	0.57	0.047
Lit. β_{11}	-0.46	0.465	0.51	0.278	-0.72	0.059	-0.32	0.333	-0.34	0.224	0.59	0.004
Math β_{11}	-0.05	0.887	-0.37	0.366	-0.73	0.030	-0.26	0.400	-0.32	0.219	0.69	0.007

Note. The number of female children for each age band cohort is B-1 (n = 1,463, 48.8%); 1-2 n = (1,447, 48.2%); 2-3 n = (1,420, 47.3%); preschool (n = 1,443, 48.1%); pre-k (n = 1423, 47.4%); and kindergarten (n = 1,459, 48.6%).

Table 11
Female Effect on Age Growth Parameters

	Birth to One		One to Two		Two to Three		Preschool		Pre-K		Kindergarten	
	β	p	β	p	β	p	β	p	β	p	β	p
Intercept, π_2												
Social β_{21}	-2.36	0.447	0.33	0.910	10.05	<0.001	13.98	<0.001	13.65	<0.001	11.52	<0.001
Phys. β_{21}	0.18	0.967	-1.13	0.768	10.20	0.013	6.33	0.182	12.01	0.007	7.88	0.117
Lang. β_{21}	1.61	0.623	5.38	0.095	17.72	<0.001	25.77	<0.001	16.58	<0.001	14.01	<0.001
Cog. β_{21}	1.76	0.590	-2.23	0.437	10.22	<0.001	13.23	<0.001	10.07	0.008	10.75	0.002
Lit. β_{21}	6.89	0.276	2.35	0.627	14.28	<0.001	19.86	<0.001	11.53	<0.001	5.54	0.034
Math β_{21}	6.19	0.072	8.80	0.026	12.76	<0.001	16.29	<0.001	9.22	0.004	2.25	0.506
Slope, π_3												
Social β_{31}	0.53	0.057	0.38	0.161	0.07	0.809	0.23	0.428	-0.06	0.841	0.61	0.017
Phys. β_{31}	0.29	0.444	0.30	0.405	-0.41	0.279	0.35	0.345	-0.30	0.413	0.02	0.962
Lang. β_{31}	0.43	0.145	0.34	0.263	-0.32	0.344	-0.27	0.445	-0.07	0.835	0.43	0.144
Cog. β_{31}	0.21	0.461	0.62	0.025	-0.17	0.557	0.14	0.656	0.29	0.356	0.49	0.066
Lit. β_{31}	-0.13	0.824	0.48	0.270	-0.50	0.161	-0.44	0.130	-0.39	0.142	0.50	0.011
Math β_{31}	-0.10	0.786	-0.34	0.357	-0.58	0.061	-0.39	0.161	-0.31	0.206	0.56	0.021

Note. The number of female children for each age band cohort is B-1 (n = 1,463, 48.8%); 1-2 n = (1,447, 48.2%); 2-3 n = (1,420, 47.3%); preschool (n = 1,443, 48.1%); pre-k (n = 1423, 47.4%); and kindergarten (n = 1,459, 48.6%).

Spanish-Speaking Children

Research suggests that children whose primary language is Spanish demonstrate different patterns of growth than children whose primary language is English (Hujar et al., 2021; Roberts & Bryant, 2011). Research tends to suggest that speaking Spanish is associated with lower preliminary abilities across academic and developmental domains (Hujar et al., 2021; Roberts & Bryant, 2011) and stronger rates of growth in early childhood (Hujar et al., 2021; Roberts & Bryant, 2011). Given previous research, it was hypothesized that children whose primary language was Spanish would demonstrate lower initial levels of development and stronger growth rates over time across domains of learning and development.

Despite extant literature, which suggests that children who speak Spanish tend to demonstrate lower preliminary levels of development across domains (Hujar et al., 2021; Roberts & Bryant, 2011), results from the present study suggested children who speak Spanish only demonstrated lower patterns of performance in some domains of learning and development. To begin, results including negative intercept beta coefficients, β_{02}, and statistically significant p-values suggested that Spanish-speaking children in one- to two-year-old classrooms and beyond demonstrated lower initial language development levels than children whose primary language was English. For example, Spanish-speaking children in one- to two-year-old classrooms earned on average 17.97 (p =.002) fewer language scale points than their English-speaking peers prior to instruction. This trend persisted for Spanish-speaking children in two- to three-year-old classrooms (β_{02}= -15.08, p = .027), preschool classrooms (β_{02}= -22.35, p =.001), pre-k classrooms (β_{02}= -27.49, p =.001), and kindergarten classrooms (β_{02} = -17.24, p =.004). Similarly, negative intercept beta coefficients, β_{22}, and statistically significant p-values suggested that Spanish-speaking children in one- to two-year-old, two- to three-year-old, preschool, and pre-k classrooms typically demonstrated lower initial language development levels than similarly aged children whose primary language was English. For example, Spanish-speaking children in one- to two-year-old classrooms earned on average 20.30 (p = .004) fewer language scale points at 12 months of age, when compared to their English-speaking peers. This trend persisted for Spanish-speaking children in two- to three-year-old classrooms (β_{22} = -17.31, p = .039), preschool classrooms (-34.11, p <.001), and pre-k classrooms (β_{22}= -25.19, p .009).

Additionally, older children whose primary language was Spanish, including children in pre-k and kindergarten classrooms demonstrated lower preliminary academic abilities when compared to children whose primary language was English. Negative intercept beta coefficients, β_{02} and β_{22}, and statistically significant p-values suggested that children in pre-k and kindergarten experienced additional adverse effects in academic domains, including mathematics and literacy. For example, Spanish-speaking children in pre-k earned on average, 14.45 (p = .005) mathematics and 14.20 (p = .005) literacy scale points less prior to instruction and Spanish-speaking kindergarteners earned on average, 29.32 (p < .001) mathematics and 13.91 (p < .001) literacy scale points less prior to instruction than children whose primary language was English. Collectively, these results suggest that speaking Spanish may present additional challenges for children as academic rigor increases and content-specific vocabulary becomes more complex.

Finally, despite previous research studies, which have found that Spanish-speaking children tend to demonstrate stronger rates of growth over time when compared to their English-speaking peers (Hujar et al., 2021; Roberts & Bryant, 2011), results from the present study did not demonstrate any consistent patterns of differences in growth rates across age bands, domains, and time metrics. However, Spanish-speaking preschoolers and kindergarteners demonstrated stronger growth in some areas of learning and development. Positive language and cognitive slope coefficients, β_{12} and β_{32}, suggested that preschoolers experienced stronger growth rates across time metrics. For example, Spanish-speaking preschoolers gained on average 2.15 (p = .008) additional language scale points and 1.40 (p = .040) additional cognitive scale points per month of instruction when compared to their English-speaking peers. Similarly, positive mathematics slope coefficients, β_{12} and β_{32}, suggested that kindergarteners experienced stronger growth rates across months of instruction and months of age. For example, Spanish-speaking kindergarteners gained on average 1.48 (p = .009) additional mathematics scale points per month of instruction than their English-speaking peers. Together these results suggest that while inconsistent, some Spanish-speaking children may experience some type of catch-up effect. See Tables 12 and 13 for Spanish effects on instructional growth parameters and age growth parameters respectively. Positive and statistically significant beta coefficients are highlighted in green to emphasize favorable effects. Conversely, negative and statistically significant beta coefficients are highlighted in red to emphasize adverse effects.

Table 12

Spanish Effect on Instructional Growth Parameters

	Birth to One		One to Two		Two to Three		Preschool		Pre-K		Kindergarten	
	β	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*
Intercept, π_0												
Social β_{02}	-9.04	0.062	-14.76	<0.001	-7.22	0.150	-2.87	0.562	-7.18	0.140	-3.79	0.409
Phys. β_{02}	-8.98	0.161	-16.55	0.005	-3.13	0.623	-4.75	0.467	-9.90	0.138	-9.32	0.258
Lang. β_{02}	-0.49	0.924	-17.97	0.002	-15.08	0.027	-22.35	0.001	-27.49	<0.001	-17.24	0.004
Cog. β_{02}	-6.16	0.205	-14.05	0.002	-6.06	0.230	-3.86	0.468	-9.85	0.065	-8.17	0.071
Lit. β_{02}	6.57	0.498	-21.92	0.012	-21.34	0.006	-6.58	0.251	-14.20	0.005	-13.91	<0.001
Math β_{02}	0.10	0.980	-5.21	0.410	-8.58	0.152	0.56	0.915	-14.45	0.003	-29.32	<0.001
Slope, π_1												
Social β_{12}	0.29	0.659	0.870	0.122	0.84	0.184	1.10	0.091	0.70	0.275	-0.55	0.298
Phys. β_{12}	0.03	0.974	0.42	0.570	-0.17	0.840	1.15	0.175	1.21	0.153	0.29	0.728
Lang. β_{12}	-0.71	0.304	0.75	0.284	0.30	0.714	2.15	0.008	0.70	0.408	-0.79	0.213
Cog. β_{12}	0.24	0.689	1.00	0.090	0.70	0.285	1.40	0.040	1.10	0.102	-0.87	0.102
Lit. β_{12}	-1.10	0.407	0.75	0.469	1.60	0.087	1.05	0.102	0.96	0.115	-0.30	0.464
Math β_{12}	1.27	0.084	-0.15	0.857	1.38	0.061	0.21	0.716	0.76	0.178	1.48	0.009

Note. The number of children whose primary language is Spanish for each age band cohort is B-1 (*n* = 417, 13.9%); 1-2 (*n* = 397, 13.2%); 2-3 (*n* =436, 14.5%); preschool (*n* = 404, 13.5%); pre-k (*n* = 384, 12.8%); and kindergarten (*n* = 416, 13.9%).

Table 13

Spanish Effect on Age Growth Parameters

	Birth to One		One to Two		Two to Three		Preschool		Pre-K		Kindergarten	
	β	p	β	p	β	p	β	p	β	p	β	p
Intercept, π_2												
Social β_{22}	-9.18	0.159	-15.95	0.004	-7.82	0.215	-10.90	0.108	-4.65	0.504	-0.12	0.985
Phys. β_{22}	-8.75	0.282	-18.44	0.013	-1.74	0.829	-13.73	0.129	-9.45	0.301	-14.91	0.162
Lang. β_{22}	4.95	0.478	-20.30	0.004	-17.31	0.039	-34.11	<0.001	-25.19	0.009	-13.51	0.092
Cog. β_{22}	-4.64	0.472	-17.85	0.001	-5.74	0.363	-12.80	0.076	-6.94	0.362	-5.44	0.403
Lit. β_{22}	16.93	0.218	-20.57	0.067	-23.60	0.016	-13.65	0.077	-13.31	0.074	-13.16	0.013
Math β_{22}	-1.43	0.818	-4.40	0.582	-8.26	0.266	-3.92	0.575	-11.47	0.098	-33.23	<0.001
Slope, π_3												
Social β_{32}	0.08	0.890	0.36	0.481	0.48	0.407	1.09	0.048	0.31	0.591	-0.69	0.164
Phys. β_{32}	-0.13	0.860	0.07	0.92	-0.40	0.590	1.11	0.127	0.91	0.217	0.61	0.451
Lang. β_{32}	-0.92	0.147	0.44	0.493	0.29	0.703	1.87	0.010	0.21	0.795	-0.85	0.154
Cog. β_{32}	-0.11	0.838	0.69	0.192	0.27	0.655	1.32	0.027	0.51	0.410	-0.83	0.107
Lit. β_{32}	-1.70	0.159	0.02	0.980	1.15	0.192	0.93	0.097	0.51	0.398	-0.29	0.466
Math β_{32}	0.79	0.234	-0.65	0.380	0.72	0.290	0.26	0.626	0.31	0.551	1.14	0.037

Note. The number of children whose primary language is Spanish for each age band cohort is B-1 (n = 417, 13.9%); 1-2 (n = 397, 13.2%); 2-3 (n =436, 14.5%); preschool (n = 404, 13.5%); pre-k (n = 384, 12.8%); and kindergarten (n = 416, 13.9%).

Children who Speak Other Languages

Researchers have found that conducting subgroup analyses is challenging because researchers must create homogenous groups (Roberts & Bryant, 2011). Yet, in reality, intersecting identities and unmeasured variables, such as language proficiency, socioeconomic status, and maternal education, threaten the validity and interpretability of results. While children in this group are similar to one another in the sense that they speak languages other than English or Spanish, the reality is, children in this group are also very different from one another. Across the full sample, children in this group speak 53 unique languages, and each age-band cohort includes children who speak 20-35 different languages. Given within-group and between-cohort heterogeneity, there were no a priori hypotheses regarding differences in preliminary performance and patterns of growth over time for children in this group.

While there were no consistent differences in preliminary performance across age bands and domains, two- and three-year-old children and kindergarteners whose primary language was *other* generally demonstrated lower preliminary developmental levels when compared to their English-speaking peers. Negative intercept beta coefficients, β_{03} and β_{23}, and statistically significant *p*-values suggested that two- to three-year-old children whose primary language was not English or Spanish demonstrated lower initial levels of development across every domain of development, when compared to their English-speaking peers. For example, children in two- to three-year-old classrooms whose primary language was categorized as *other* earned on average 16.38 (p = .002) fewer literacy scale points prior to instruction than their English-speaking peers. Similarly, kindergarteners whose primary language was not English or Spanish also demonstrated lower preliminary levels of physical, language, cognitive, and literacy development, when compared to their English-speaking peers. For example, kindergarteners whose primary language was categorized as *other* earned on average 5.40 (*p* = .011) fewer literacy scale points than their English-speaking peers prior to instruction.

Additionally, while there were not any consistent differences in growth rate over time across age bands and domains, children in two- to three-year-old, preschool, and kindergarten classrooms demonstrated stronger growth rates across many domains of learning and development. Positive slope beta coefficients, β_{13} and β_{33}, and statistically significant *p*-values suggested that two- to three-year-old children whose primary language was not English or Spanish demonstrated stronger social-emotional, physical, cognitive, literacy and mathematics growth. For example, children in two- to three-year-old classrooms whose primary language was identified as *other*, gained an additional 1.81 (*p* = .008) literacy scale points per month of instruction than their English-speaking peers. Similarly, positive slope beta coefficients, β_{13} and β_{33}, and statistically significant *p*-values suggested that preschool children whose primary language was identified as *other* also experienced stronger physical, language, cognitive, literacy, and mathematics growth than their English-speaking peers. For example, preschoolers whose primary language was not English or Spanish, gained an additional 1.81 (*p* = .012) literacy scale points per month of instruction. And finally, positive slope beta coefficients, β_{13} and β_{33}, and statistically significant *p*-values suggested that kindergarteners whose primary language was identified as *other* experienced stronger physical, language, cognitive, and literacy growth. For example, kindergarteners whose primary language was identified as *other*, gained on average 1.07 (*p* < .001) additional literacy scale points per month of instruction when compared to children whose primary language was English. See Tables 14 and 15 for *other* language effects on instructional growth parameters and age growth parameters respectively. Positive and statistically significant beta coefficients are highlighted in green to emphasize favorable effects.

Conversely, negative and statistically significant beta coefficients are highlighted in red to emphasize adverse effects.

Table 14

Other Language Effect on Instructional Growth Parameters

	Birth to One		One to Two		Two to Three		Preschool		Pre-K		Kindergarten	
	β	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*
Intercept, π_0												
Social β_{03}	-5.76	0.174	4.73	0.296	-17.94	<0.001	-4.43	0.350	-5.53	0.228	-10.35	<0.001
Phys. β_{03}	-9.17	0.144	-1.06	0.867	-27.24	<0.001	-8.63	0.215	-7.04	0.314	-10.53	0.019
Lang. β_{03}	-4.58	0.305	6.73	0.203	-17.18	0.001	-13.74	0.035	-8.95	0.154	-17.37	<0.001
Cog. β_{03}	-3.18	0.485	4.32	0.377	-17.35	<0.001	-8.36	0.121	-4.66	0.431	-12.05	<0.001
Lit. β_{03}	-9.98	0.280	8.30	0.256	-16.38	0.002	-7.51	0.153	-6.15	0.270	-5.40	0.011
Math β_{03}	-4.35	0.310	3.16	0.633	-19.98	<0.001	-8.63	0.090	-6.24	0.208	0.35	0.895
Slope, π_1												
Social β_{13}	-0.53	0.368	-0.623	0.252	1.52	0.013	1.11	0.092	-0.07	0.92	2.34	<0.001
Phys. β_{13}	0.47	0.574	-0.16	0.840	2.47	<0.001	2.35	0.004	0.14	0.866	2.95	<0.001
Lang. β_{13}	-0.33	0.558	-0.80	0.194	1.16	0.081	2.29	0.006	-0.83	0.285	2.40	<0.001
Cog. β_{13}	-0.77	0.147	-0.69	0.235	1.52	0.016	1.85	0.015	0.21	0.763	2.32	<0.001
Lit. β_{13}	1.06	0.370	-0.83	0.331	1.81	0.008	1.81	0.012	0.24	0.701	1.07	<0.001
Math β_{13}	0.47	0.520	-0.14	0.872	1.91	0.003	1.62	0.007	0.26	0.616	0.28	0.342

Note. The number of children who speak a primary language that is not English or Spanish for each age band cohort is B-1 (*n* = 250, 8.3%); 1-2 (*n* = 292, 9.7%); 2-3 (*n* =290, 9.7%); preschool (*n* = 309, 10.3%); pre-k (*n* = 250, 8.3%); and kindergarten (*n* = 726, 24.2%).

Table 15

Other Language Effect on Age Growth Parameters

	Birth to One		One to Two		Two to Three		Preschool		Pre-K		Kindergarten	
	β	p	β	p	β	p	β	p	β	p	β	p
Intercept, π_2												
Social β_{23}	-0.69	0.905	6.46	0.231	-15.98	0.007	-4.63	0.497	-6.52	0.306	-16.98	<0.001
Phys. β_{23}	-6.18	0.453	1.87	0.813	-24.94	<0.001	-11.68	0.195	-11.30	0.207	-19.61	0.001
Lang. β_{23}	-1.52	0.790	8.19	0.184	-15.22	0.018	-20.39	0.019	-8.13	0.313	-24.70	<0.001
Cog. β_{23}	2.90	0.633	5.44	0.353	-15.29	0.009	-10.40	0.175	-8.48	0.262	-19.19	<0.001
Lit. β_{23}	-6.57	0.618	10.75	0.229	-13.83	0.034	-10.75	0.148	-8.85	0.227	-7.04	0.020
Math β_{23}	-0.48	0.946	1.32	0.867	-17.19	0.009	-12.36	0.066	-10.20	0.109	1.08	0.777
Slope, π_3												
Social β_{33}	-0.84	0.119	-0.64	0.19	0.93	0.109	0.60	0.343	-0.06	0.921	1.93	<0.001
Phys. β_{33}	-0.20	0.797	-0.47	0.525	1.58	0.018	1.54	0.037	0.37	0.624	2.48	<0.001
Lang. β_{33}	-0.52	0.332	-0.75	0.189	0.77	0.218	1.89	0.012	-0.63	0.386	2.02	<0.001
Cog. β_{33}	-1.09	0.045	-0.64	0.227	0.97	0.096	1.20	0.084	0.32	0.624	1.98	<0.001
Lit. β_{33}	0.23	0.840	-0.87	0.260	1.08	0.074	1.28	0.046	0.34	0.55	0.75	<0.001
Math β_{33}	-0.13	0.849	0.02	0.978	1.17	0.045	1.21	0.024	0.41	0.402	0.06	0.827

Note. The number of children who speak a primary language that is not English or Spanish for each age band cohort is B-1 (n = 250, 8.3%); 1-2 (n = 292, 9.7%); 2-3 (n =290, 9.7%); preschool (n = 309, 10.3%); pre-k (n = 250, 8.3%); and kindergarten (n = 726, 24.2%).

Black Non-Hispanic Children

Research suggests that Black Non-Hispanic children tend to demonstrate lower abilities early on (Kuhfeld et al., 2020; Readon & Portilla, 2016) that persist through high school (Young et al., 2017). Therefore, it was hypothesized that Black Non-Hispanic children would demonstrate lower patterns of preliminary performance and growth over time. While results, including negative intercept beta coefficients, β_{04} and , β_{24}, and statistically significant p-values generally suggested that Black Non-Hispanic children in one- to two-year-old, two- to three-year-old, preschool, pre-k, and kindergarten classrooms demonstrated lower preliminary levels of development when compared to their White Non-Hispanic peers across many domains of learning and development, positive intercept beta coefficients, β_{04} and , β_{24}, and statistically significant p-values generally suggested Black Non-Hispanic children in birth- to one-year-old classrooms demonstrated higher preliminary levels of development when compared to their White Non-Hispanic peers. While it's possible that Black Non-Hispanic children in birth- to one-year-old classrooms experience an advantage over their White Non-Hispanic peers in the areas of language, cognitive, literacy, and mathematics development, it's also possible that unmeasured variables confounded results.

Finally, despite previous studies, which have found that Black Non-Hispanic children tend to demonstrate slower patterns of growth than their White Non-Hispanic peers (Reardon & Portilla, 2016; Young et al., 2017), Black Non-Hispanic children in the current study did not demonstrate consistently different patterns of growth over time across age bands, domains, and time metrics. Positive slope beta coefficients, β_{14} and β_{34}, suggested that Black Non-Hispanic preschoolers experienced stronger social-emotional, physical, and cognitive growth for each month of instruction or age. For example, Black Non-Hispanic preschoolers gained on average, an additional 1.91 (*p* = .010) cognitive scale points for each month of instruction and an additional 1.49 (*p* = .012) cognitive scale points for each month of age. Conversely, Black Non-Hispanic children in birth- to one-year-old classrooms and kindergarten classrooms experienced slower growth rates in several areas of learning and development when compared to their White Non-Hispanic peers. For example, Black Non-Hispanic children in birth- to one-year-old classrooms gained on average, 1.01 (*p* = .036) fewer cognitive scale points and Black Non-Hispanic kindergarteners gained on average 1.64 (*p* = .002) fewer cognitive scale points per month of instruction, when compared to their White Non-Hispanic peers. See Tables 16 and 17 for Black Non-Hispanic effects on instructional growth parameters and age growth parameters respectively. Positive and statistically significant beta coefficients are highlighted in green to emphasize favorable effects. Conversely, negative and statistically significant beta coefficients are highlighted in red to emphasize adverse effects.

Table 16

Black Non-Hispanic Effect on Instructional Growth Parameters

	Birth to One		One to Two		Two to Three		Preschool		Pre-K		Kindergarten	
	β	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*
Intercept, π_0												
Social β_{04}	2.11	0.603	-6.81	0.066	-9.99	0.009	-1.99	0.622	-10.32	0.004	-3.09	0.475
Phys. β_{04}	2.40	0.679	-11.92	0.016	-20.71	<0.001	-10.96	0.057	-11.46	0.021	-14.54	0.040
Lang. β_{04}	2.92	0.491	-11.29	0.010	-25.73	<0.001	-4.19	0.426	-15.32	<0.001	-4.90	0.365
Cog. β_{04}	7.97	0.059	-9.77	0.014	-18.59	<0.001	-11.23	0.027	-15.08	<0.001	-10.32	0.030
Lit. β_{04}	18.36	0.020	-6.86	0.273	-13.46	0.008	0.80	0.851	-3.19	0.317	-8.25	0.018
Math β_{04}	16.19	<0.001	-6.77	0.221	-17.47	<0.001	-0.70	0.874	-8.26	0.017	-9.90	0.008
Slope, π_1												
Social β_{14}	-0.47	0.362	-0.124	0.812	-0.07	0.887	1.56	0.009	1.34	0.009	-0.96	0.060
Phys. β_{14}	-0.83	0.221	-0.06	0.928	-0.01	0.986	2.19	0.003	1.15	0.066	-0.19	0.776
Lang. β_{14}	-0.84	0.107	-0.68	0.232	0.02	0.971	1.39	0.057	1.01	0.102	-2.22	<0.001
Cog. β_{14}	-1.01	0.036	-0.23	0.677	-0.30	0.548	1.91	0.010	1.07	0.062	-1.64	0.002
Lit. β_{14}	-2.03	0.050	0.24	0.769	0.64	0.266	0.93	0.081	-0.06	0.890	-1.15	0.001
Math β_{14}	-0.26	0.667	-0.10	0.886	0.05	0.926	0.93	0.112	0.61	0.146	-1.47	<0.001

Note. The number of Black Non-Hispanic children for each age band cohort is B-1 (*n* = 421, 14.0%); 1-2, 2-3, preschool, and pre-k (*n* = 412, 13.7%); and kindergarten (*n* = 415, 13.8%).

Table 17

Black Non-Hispanic Effect on Age Growth Parameters

	Birth to One		One to Two		Two to Three		Preschool		Pre-K		Kindergarten	
	β	p	β	p	β	p	β	p	β	p	β	p
Intercept, π_2												
Social β_{24}	9.07	0.101	-8.30	0.078	-12.38	0.013	-7.14	0.239	-14.02	0.012	0.13	0.983
Phys. β_{24}	14.80	0.057	-14.29	0.022	-23.39	<0.001	-16.88	0.038	-15.31	0.031	-12.15	0.189
Lang. β_{24}	11.38	0.049	-8.91	0.098	-27.83	<0.001	-9.21	0.222	-17.88	0.007	1.69	0.814
Cog. β_{24}	17.61	0.002	-8.69	0.078	-20.81	<0.001	-17.94	0.014	-18.66	0.002	-5.27	0.420
Lit. β_{24}	38.12	<0.001	-5.21	0.512	-18.01	0.006	-2.20	0.720	-1.70	0.726	-4.77	0.325
Math β_{24}	24.18	<0.001	-5.20	0.436	-20.10	<0.001	-2.69	0.670	-11.23	0.025	-5.29	0.312
Slope, π_3												
Social β_{34}	-0.96	0.051	0.14	0.758	0.14	0.772	1.18	0.022	0.87	0.069	-0.84	0.069
Phys. β_{34}	-1.65	0.014	0.33	0.586	0.17	0.788	1.52	0.016	0.75	0.190	-0.37	0.560
Lang. β_{34}	-1.33	0.009	-0.60	0.245	0.11	0.835	1.06	0.087	0.56	0.328	-1.86	<0.001
Cog. β_{34}	-1.56	0.001	-0.19	0.693	-0.05	0.916	1.49	0.012	0.70	0.180	-1.41	0.004
Lit. β_{34}	-3.13	0.001	0.07	0.919	0.82	0.134	0.68	0.140	-0.33	0.447	-0.98	0.004
Math β_{34}	-1.11	0.069	-0.12	0.848	0.21	0.695	0.57	0.250	0.42	0.306	-1.27	<0.001

Note. The number of Black Non-Hispanic children for each age band cohort is B-1 (n = 421, 14.0%); 1-2, 2-3, preschool, and pre-k (n = 412, 13.7%); and kindergarten (n = 415, 13.8%).

Hispanic Children

Research suggests that Hispanic children tend to demonstrate lower preliminary abilities at school entry (Kuhfeld et al., 2020; Readon & Portilla, 2016) and stronger growth in the first few years of school (Hujar et al., 2021; Reardon & Galindo, 2009). Given previous research, it was hypothesized that Hispanic children would demonstrate lower preliminary levels of development and stronger growth rates over time. Results generally suggested that Hispanic children in two- to three-year-old classrooms and beyond demonstrated lower preliminary developmental levels than their White Non-Hispanic counterparts. Negative intercept beta coefficients, β_{08} and β_{28}, and statistically significant p-values suggested that Hispanic children in two- to three-year-old, preschool, pre-k, and kindergarten classrooms typically demonstrated lower preliminary levels of development when compared to their White Non-Hispanic peers. For example, Hispanic children in two- to three-year-old classrooms earned on average, 15.42 (p = .002) mathematics scale points less than their White Non-Hispanic peers prior to instruction. This pattern persisted through preschool (β_{08} = -10.12, p = .025), pre-k (β_{08} = -14.85, p <.001), and kindergarten (β_{08} = -22.03, p <.001). Similarly, Hispanic children in two- to three-year-old classrooms who were 24 months of age at the beginning of the academic year earned on average, 14.45 (p = .002) fewer mathematics scale points. This trend persisted for Hispanic children in pre-k (β_{28} = -18.78, p <.001) and kindergarten (β_{28} = -25.44, p <.001).

Despite previous studies, which have found that Hispanic children experience accelerated growth during the first few years of school (Hujar et al., 2021; Roberts & Bryant, 2011), Hispanic children in the current study did not demonstrate significantly stronger growth rates when compared to their White Non-Hispanic peers. In fact, Hispanic kindergarteners actually experienced slower social-emotional and physical growth than their White Non-Hispanic peers.

For example, Hispanic kindergarteners on average gained 1.03 (*p* = .023) social-emotional scale points less per month of instruction than their White counterparts. While these observable differences in growth rate could be due to true differences in patterns of performance, it’s also possible that unmeasured variables such as socioeconomic status could have influenced results. See Tables 18 and 19 for Hispanic effects on instructional growth parameters and age growth parameters respectively. Positive and statistically significant beta coefficients are highlighted in green to emphasize favorable effects. Conversely, negative and statistically significant beta coefficients are highlighted in red to emphasize adverse effects.

Table 18

Hispanic Effect on Instructional Growth Parameters

	Birth to One		One to Two		Two to Three		Preschool		Pre-K		Kindergarten	
	β	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*
Intercept, π_0												
Social β_{08}	-1.81	0.649	-4.89	0.185	-12.82	0.003	-5.93	0.157	-9.31	0.013	-1.60	0.661
Phys. β_{08}	-0.90	0.864	-9.08	0.059	-22.95	<0.001	-8.02	0.126	-9.74	0.042	-4.93	0.398
Lang. β_{08}	-0.64	0.867	-11.43	0.005	-22.49	<0.001	-18.89	<0.001	-20.09	<0.001	-15.42	<0.001
Cog. β_{08}	4.47	0.260	-7.16	0.054	-14.84	<0.001	-9.76	0.031	-13.61	<0.001	-10.58	0.004
Lit. β_{08}	8.26	0.247	-3.41	0.563	-13.46	0.014	-9.45	0.039	-10.89	0.001	-14.45	<0.001
Math β_{08}	3.49	0.299	-6.13	0.227	-15.42	0.002	-10.12	0.025	-14.85	<0.001	-22.03	<0.001
Slope, π_1												
Social β_{18}	-0.15	0.773	-0.457	0.312	0.02	0.973	0.25	0.647	0.67	0.169	-1.03	0.023
Phys. β_{18}	-0.01	0.987	-0.31	0.610	0.37	0.608	0.62	0.357	0.51	0.392	-1.00	0.101
Lang. β_{18}	-0.19	0.700	-0.75	0.133	-0.40	0.534	0.29	0.642	0.52	0.380	-0.59	0.257
Cog. β_{18}	-0.86	0.074	-0.45	0.340	-0.15	0.798	0.09	0.873	0.39	0.434	-0.52	0.232
Lit. β_{18}	-1.29	0.179	-1.11	0.115	-0.73	0.296	0.52	0.308	0.24	0.524	-0.33	0.350
Math β_{18}	-0.26	0.667	-0.66	0.315	-1.04	0.095	0.35	0.473	0.73	0.060	0.70	0.115

Note. The number of Hispanic children for each age band cohort is B-1 (*n* = 783, 26.1%); 1-2, 2-3, preschool, and pre-k (*n* = 767, 25.6%); and kindergarten (*n* = 773, 25.8%).

Table 19

Hispanic Effect on Age Growth Parameters

	Birth to One		One to Two		Two to Three		Preschool		Pre-K		Kindergarten	
	β	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*	*β*	*p*
Intercept, π_2												
Social β_{28}	-1.30	0.803	-2.10	0.640	-13.93	0.009	-3.27	0.567	-12.23	0.025	1.45	0.788
Phys. β_{28}	1.72	0.793	-6.32	0.288	-22.59	0.001	-5.35	0.453	-10.71	0.109	1.93	0.798
Lang. β_{28}	-0.42	0.935	-4.11	0.398	-20.72	0.002	-15.44	0.034	-22.19	0.001	-14.77	0.022
Cog. β_{28}	6.57	0.196	-2.54	0.573	-15.33	0.005	-7.72	0.203	-16.07	0.004	-10.13	0.061
Lit. β_{28}	12.96	0.170	4.32	0.569	-11.86	0.083	-8.79	0.152	-11.55	0.018	-13.94	0.003
Math β_{28}	4.28	0.378	0.94	0.880	-14.45	0.020	-8.03	0.184	-18.78	<0.001	-25.44	<0.001
Slope, π_3												
Social β_{38}	-0.23	0.623	-0.34	0.411	0.25	0.616	0.09	0.845	0.59	0.197	-0.85	0.046
Phys. β_{38}	-0.46	0.446	-0.16	0.784	0.33	0.615	0.30	0.605	0.22	0.686	-1.21	0.048
Lang. β_{38}	-0.22	0.638	-0.94	0.043	-0.32	0.594	0.10	0.856	0.46	0.397	-0.376	0.442
Cog. β_{38}	-0.80	0.083	-0.50	0.242	0.09	0.864	0.08	0.879	0.38	0.414	-0.34	0.411
Lit. β_{38}	-1.35	0.109	-1.21	0.070	-0.54	0.401	0.41	0.360	0.16	0.664	-0.24	0.494
Math β_{38}	-0.27	0.616	-0.77	0.199	-0.64	0.261	0.20	0.661	0.71	0.056	0.73	0.091

Note. The number of Hispanic children for each age band cohort is B-1 (*n* = 783, 26.1%); 1-2, 2-3, preschool, and pre-k (*n* = 767, 25.6%); and kindergartern (*n* = 773, 25.8%).

CHAPTER FIVE: CONCLUSIONS

As discussed in Chapter One, the primary purpose of this study was to provide more precise and time-continuous normative ability and growth scores that were reflective of the intra-individual growth process and leveraged between-child variability in age and instructional exposure (Pan & Goldstein, 1997; Thum & Kuhfeld, 2020). Secondary study purposes included a) providing evidence of the most effective time metric, age or instructional exposure, for modeling developmental growth from birth through kindergarten, b) establishing evidence of the nature of the growth process from birth through kindergarten, and c) examining the relationships between notable child-level characteristics including, gender, race/ethnicity, and primary language, and growth trajectories across domains of learning and development.

To lay the foundation for the present study, Chapter Two included a thorough review of extant literature on the topics of early childhood assessment, developmental growth trends, and approaches to modeling growth over time. Next, Chapter Three reviewed the methods, including the measure, population, sampling plan, and data collection and analysis procedures. Chapter Four included the results corresponding to each research question. And finally, Chapter Five includes key results and implications and recommendations for both research and practice.

Model-Estimated Normative Scores

The first research question aimed to understand typical development at each point in time from birth through kindergarten for every major domain of learning and development presented in the *GOLD*® assessment system. Unconditional hierarchical linear growth models were used to estimate growth parameters, including intercepts and slopes. Model-estimated intercepts and slopes were used to derive normative ability estimates for every domain, age band, and month of instruction or three months of age. While this research question can be answered by examining the Instructional *GOLD*® Normative Scores provided in Table 1 and the Age *GOLD*® Normative Scores provided in Table 4, it's also important to understand whether national normative estimates are reasonably valid.

Once again, validity "refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA et al., 2014, p. 11). However, within the context of the current study, validity is used to describe the stability and accuracy of model-estimated parameters and subsequent national normative estimates. While valid and reliable normative estimates alone do not fully substantiate the norm-referenced validity argument, they are prerequisite to making appropriate, useful, and meaningful relative score interpretations. Therefore, the validity of model-based estimates and corresponding national normative scores was examined using several strategies. First, the reliability of model-estimated intercepts was examined. Results suggested that model-estimated intercepts were reasonably reliable across age bands for Unconditional Instructional Exposure (.81-.93) and Unconditional Age (.71-.94) Growth Models. Second, we would expect normative estimates to increase across age bands and months of age and instruction. Results, including model-estimated normative scores presented in Table 1 and Table 4 evidenced this trend. Third, for each domain and age band, the Unconditional Instructional Exposure Model-estimated should be higher than the Unconditional Age in Months Model-estimated intercept. Results, including model-estimated intercepts consistently demonstrated this trend across age bands and domains of learning and development. And finally, HLM-estimated instructional normative scores were compared to score distribution-based norms provided in the *Technical Manual for the Teaching Strategies GOLD*® Assessment (2nd Edition) to determine if estimates were reasonably accurate. Results suggested the HLM-estimated instructional normative scores corresponding to the most common months for finalizing scores and score distribution-based norms for the average fall, winter, and spring assessment occasions provided similar and at times equivalent estimates. Together, these four sources of validity evidence form the foundation of the validity argument for the use of HLM-derived national normative *GOLD*® scores to understand relative performance for children ages birth through kindergarten.

Implications

The validity evidence presented in Chapter Four and reviewed in the present chapter has two significant implications. First, collectively, the validity evidence suggests that HLM-estimated normative scores are reasonably reliable and valid. Therefore, HLM-based estimates could be used to help teachers draw more precise conclusions about relative performance and growth over time. Compared to previous norms, which were provided for the average fall, winter, and spring assessment date for each age band, HLM-based norms provide the expected score for every month of instruction and every three months of age. By increasing the precision in time metric, teachers who finalize scores earlier or later than average or have children who are younger or older than average can still draw meaningful conclusions about relative performance and growth. Second, while researchers have used MLMs to estimate normative ability and growth scores (Thum & Kuhfeld, 2020; Pan & Goldstein, 1997), it's not a widely used method for establishing norms. Results, including validity evidence, from the present study suggest that HLMs can be used to establish normative ability and growth scores when data include at least three measurements per individual and data are collected over a relatively short period of time.

Recommendations

Results from the present study generally suggest that HLM-based normative estimates are reasonably reliable and valid. Therefore, Teaching Strategies® should consider adopting the new national normative scores to aid teachers in making more precise relative score interpretations at each point in time from birth through kindergarten. Additionally, given "that the validation process never ends, as there is always additional information that can be gathered to more fully understand a test and the inferences that can be drawn from it" (AERA et al., 2014, p. 21), future research should seek to collect additional validity evidence in support of the novel normative estimates. For example, while HLM-based estimates were compared to score distribution-based estimates to examine whether estimates were reasonably similar, HLM-based estimates and score distribution-based estimates were established using different years of data and different cohorts of children. Future research could replicate the methods used by Lambert (2020) with the current sample to examine how similar estimates are across methods. By engaging in this type of research, researchers could a) establish additional validity evidence for HLM-derived national normative scores and b) provide further evidence to support the use of HLMs to establish normative ability and growth estimates.

The Most Effective Time Metric

The second research question aimed to understand which time metric, approximate instructional exposure or age in months, was the most effective predictor of developmental growth for each age band and domain of development. Model fit statistics, including AIC and *pseudo-*r^2 statistics were compared across Unconditional Instructional Exposure Growth Models and Unconditional Age in Months Growth Models. Together, smaller AIC values and larger *pseudo-*r^2 statistics provided evidence of the most effective time metric for each age band and domain of development.

Researchers have widely observed that babies' developmental growth is closely related to biological age. For example, regardless of geographic location or beliefs about caregiver interactions, babies learn to sit, crawl, walk, and talk at similar points in time (Harkness et al., 2013; Rebello-Britto et al., 2013). Yet, researchers have also found that children as young as two- or three-years old are greatly influenced by their environment (Yoshikawa et al., 2013) and interactions with caregivers and teachers (Chetty et al., 2011; Gialamas et al., 2013). Given these widely observed trends in child development, it was hypothesized that age in months may be the most effective predictor of developmental growth for babies and toddlers ages two and younger and instructional exposure may be the most effective predictor of developmental growth for children ages two and older. While results for children in one- to two-year-old, preschool, pre-k, and kindergarten classrooms were generally in the hypothesized direction across domains of learning and development, results for children in birth- to one-year-old and two- to three-year-old classrooms were less consistent.

For children in birth- to one-year-old classrooms, age in months was the most effective predictor of social-emotional and physical development, while instructional exposure was the most effective predictor of mathematics development. Contradictory fit statistics for birth- to one-year-old language, cognitive, and literacy models suggested that neither predictor was more effective than the other. Considering the literature base which clearly suggests babies' developmental growth is closely related to biological age, we must consider alternative explanations for contradictory results. Although definitive conclusions cannot be made about conflicting fit statistics or results that oppose the study hypothesis, the most likely explanation stems from the way time metrics in the present study were measured and modeled. While age in months at the time of assessment was calculated by taking the date of assessment and subtracting the child's birthdate to obtain the child's age in months at the time of assessment, months of approximate instructional exposure was in fact approximate. Without program start dates, a typical school start date (08/14/2020) was imposed for all children in the sample. Approximate instructional exposure was calculated by taking the date of assessment and subtracting the artificial school start date to obtain each child's approximate months of instruction at the time of assessment. Given that both time variables were measured linearly and there was a strong bivariate correlation between time metrics (*r* = .68), the explanatory power of each time metric was similar regardless of domain or age band and fit statistics, including *pseudo-*r^2 statistics and AICs were very similar across non-nested models.

Similarly, contradictory fit statistics for two- to three-year-old social-emotional, physical, language, cognitive, literacy, and mathematics models indicated that neither predictor was more effective than the other. Once again, while definitive conclusions cannot be made about contradictory fit statistics, results are likely attributable to strongly correlated and linearly measured time variables. Alternatively, consistently contradictory and near equivalent model fit statistics could suggest that two- to three-year-old children's development is equally influenced by both age and instruction. While this explanation is perhaps less plausible, it is at least in part supported by the literature which suggests that children younger than two- or three-years-old are more influenced by age (Harkness et al., 2013; Rebello-Britto et al., 2013) and children ages two or three and older are profoundly influenced by the environment (Yoshikawa et al., 2013) and interactions with caregivers (Gialamas et al., 2013).

Implications

Results from this research question have three important implications. First, although differences were modest, consistent results, including smaller AICs and larger *pseudo-*r^2 statistics, suggests instructional exposure was the most effective predictor for children in preschool, pre-k, and kindergarten classrooms. Collectively these results provide support for the study hypothesis and suggest that children ages three and older need access to high-quality instruction by way of high-quality early childhood care. Second, given that age and instructional exposure were strongly correlated and linearly measured, AICs and *pseudo-r^2* statistics were very similar across non-nested models. Similar and at times contradictory fit statistics made it challenging to draw definitive conclusions about the most effective time metric, especially for some age bands and domains of learning and development. Inferences about the most effective time metric could likely be strengthened in the presence of more precise instructional exposure data. And finally, while results generally suggest that instructional exposure was a more significant predictor of developmental growth for children ages three and older, *pseudo-r^2* statistics suggest that age is generally still an important predictor of developmental growth through kindergarten. Therefore, early childhood teachers may benefit from knowing their children's ages and typical patterns of development associated with each age.

Recommendations

While differences in AICs and *pseudo-r^2* statistics were modest across non-nested models, consistent results in the hypothesized direction for preschool, pre-k, and kindergarten children suggest that instruction is a more prominent predictor of development for children ages three and older. Given previous research which suggests that children as young as two or three are profoundly influenced the learning environment (Yoshikawa et al., 2013) and interactions with teachers and caregivers (Chetty et al., 2011; Gialamas et al., 2013) and results from the present study, state- and federal-education agencies should expand access to high-quality early childhood education programs for all children ages three and older.

Next, given inconsistent results for two age bands and small differences in AICs and *pseudo-r^2* statistics across all non-nested models, researchers should seek to replicate the current study with more precise instructional exposure data. Given program-specific start dates and vacation schedules, researchers could more accurately measure and model instructional exposure. By modeling instructional exposure more precisely, fit statistics for non-nested models would likely demonstrate larger discrepancies and clearer patterns may emerge across age bands and domains of development.

Finally, small *pseudo-*r^2 statistics for one- to two-year-old models and two- to three-year-old models suggested that models for these age bands included the most residual error. While residual error by definition is unexplained variance, researchers have suggested that a significant proportion of unexplained variance in domain-level scaled scores is due to rater effects (Hujar et al., 2021). To understand which time metric is the most effective predictor of developmental growth across domains and age bands, researchers may also consider eliminating assessment records for children whose teachers are not interrater reliability certified in an effort to eliminate some degree of construct irrelevant variance and obtain model fit statistics that are more reflective of trends in developmental growth.

Nature of Developmental Growth

Unconditional Age in Months Growth Model-estimated linear slopes, β_{30}, were examined sequentially for children ages birth through kindergarten and Unconditional Instructional Exposure Growth Model-estimated linear slopes, β_{10}, were examined sequentially for children in two- to three-year-old, preschool, pre-k, and kindergarten classrooms to make inferences about the nature of developmental growth across years of instruction.

Generally, research suggests that developmental growth slows as children age and move through subsequent grade levels (Harkness et al., 2013; Lee, 2010; Masonic Institute for the Developing Brain, 2021; Mok et al., 2016; Shanley, 2016; Shin et al., 2013). Therefore, it was hypothesized that across domains of development and time metrics, linear slopes would decrease in magnitude across age bands. While model-estimated linear slopes for age in months models, β_{30}, demonstrated a pattern of deceleration across the first three age bands, β_{30}, demonstrated a pattern of acceleration across the next three age bands. For example, babies in birth- to one-year-old classrooms gained 11.38 ($p < .001$) social-emotional scale points per month of age, toddlers in one- to two-year-old classrooms gained 6.93 ($p < .001$) social-emotional scale points per month of age, children in two- to three-year-old classrooms gained 6.19 ($p < .001$) social-emotional scale points per month of age, children in preschool classrooms gained 8.18 ($p < .001$) social-emotional scale points per month of age, children in pre-k classrooms gained 9.90 ($p < .001$) social-emotional scale points per month of age, and children in kindergarten classrooms gained 11.30 ($p < .001$) social-emotional scale points per month of age.

Similarly, model-estimated linear slopes for instructional exposure models, β_{10}, demonstrated a pattern of acceleration across age bands. For example, children in two- to three-year-old classrooms gained 6.53 ($p < .001$) social-emotional scale points per month of instruction, children in preschool classrooms gained 9.12 ($p < .001$) social-emotional scale points per month of instruction, children in pre-k classrooms gained 11.36 ($p < .001$) social-emotional scale points per month of instruction, and children in kindergarten classrooms gained 12.66 ($p < .001$) social-emotional scale points per month of instruction.

While results for children ages three and older generally contradicted study hypotheses and the broader literature which suggests developmental growth slows as children age and move through subsequent grade levels (Harkness et al., 2013; Lee, 2010; Masonic Institute for the Developing Brain, 2021; Mok et al., 2016; Shanley, 2016; Shin et al., 2013), results were consistent with typical gain scores provided in the *Technical Manual for the Teaching Strategies*

GOLD® *Assessment (2^{nd} Edition) Birth Through Third Grade.* Gain scores provided in the technical manual generally suggest that developmental growth decelerates across the first three age bands and accelerates across the next three age bands. While definitive conclusions about the contradictory patterns of growth cannot be made, there are three plausible explanations for the observed results. First, differences in slopes could be reflective of true differences in growth rates between age-separated cohorts of children (Singer & Willet, 2003). Second, differences in slopes could be attributable to the method used for obtaining scaled scores, which includes the use of a curvilinear function where smaller raw score gains at the tails of the distribution contribute a greater number of scale points. And finally, growth rate, as measured by *GOLD*®, could decelerate across the first three age bands due to the natural slowing of development as children age (Shanley, 2016) and strengthen across the next three age bands as instruction intensifies.

Implications

Again, while definitive conclusions cannot be made about the nature of developmental growth from birth through kindergarten using age-separated cohort data, linear slopes, β_{30}, suggest that developmental growth decelerates from ages zero to three years accelerates from three to six years. Given this pattern of growth, it may be most appropriate for researchers with access to multi-year longitudinal *GOLD*® data to fit data to two separate polynomial models, one for children ages birth through three years and one for children ages three through six years. By incorporating a time-squared or time-cubed parameter, researchers may be able to adequately capture the deceleration in growth that occurs across the first three age bands and the acceleration that occurs across the second three age bands.

Recommendations

While inferences can be made about the nature of growth for children ages birth through kindergarten using age-separated cohort data, competing explanations for differences in growth rates cannot be eliminated. To strengthen inferences, Teaching Strategies$^{®}$ should assign and maintain consistent child IDs across academic years to establish a multiyear database. Given multiyear longitudinal data, researchers could fit data to more complex models that may demonstrate better fit to the data and more accurately capture the nature of the growth process from birth through kindergarten. For example, researchers could fit assessment data for children ages birth to age three to both linear and quadratic models. Next, researchers could use LRTs to determine whether the quadratic model demonstrated significantly better fit to the data than the linear model (O'Connell & McCoach, 2008; McCoach et al., 2022). Similarly, with more than three measurements per child, researchers could model scaled scores as a function of age in months and incorporate a level-one time-varying covariate such as instructional exposure. By modeling scaled scores as a function of both age and instructional exposure, researchers could likely explain a greater proportion of variance in the outcome and increase the precision and utility of normative estimates (Shanley, 2016; Thum & Kuhfeld, 2020).

Subgroup Differences

The fourth research question sought to investigate how growth trends differ by subgroups of children. Conditional Instructional Exposure Growth Models and Conditional Age in Months Growth Models were fitted to the data to understand the unique effect of each child-level characteristics on preliminary developmental status and growth rate. Positive and statistically significant beta coefficients indicated the group experienced stronger preliminary developmental levels or stronger growth over time, when compared to the reference group (White Non-

Hispanic, English-speaking, boys). Conversely, negative and statistically significant beta coefficients suggested that the group experienced lower preliminary developmental levels or slower growth over time, when compared to the reference group.

Girls

Research generally suggests that girls demonstrate stronger abilities early on (Kuhfeld et al., 2020; Voyer & Voyer, 2014) that persist over time (Voyer & Voyer, 2014). Therefore, it was hypothesized that girls would demonstrate stronger preliminary abilities and growth rates, when compared to boys. Results including positive and statistically significant intercept beta coefficients generally suggested that girls ages two and older demonstrated stronger preliminary developmental levels across domains and time metrics, when compared to boys. Additionally, positive slope beta coefficients suggested that kindergarten girls demonstrated stronger social-emotional, cognitive, literacy, and mathematics growth across months of instruction and stronger literacy and mathematics growth for each additional month of age, when compared to boys. Together, these results suggest that gaps between boys and girls emerge around the time children are two years old and gaps between boys and girls may begin to widen around age five or six.

Children who Speak Spanish

Researchers generally agree that children whose primary language is Spanish demonstrate lower abilities early on (Hujar et al., 2021; Roberts & Bryant, 2011) and stronger growth rates throughout early childhood (Hujar et al., 2021; Roberts & Bryant, 2011). Therefore, it was hypothesized that children whose primary language was Spanish would demonstrate lower preliminary abilities and stronger growth rates, when compared to children whose primary language was English. However, results suggested that Spanish-speaking children only demonstrated lower patterns of preliminary performance in certain areas of development. Results including negative and statistically significant intercept beta coefficients suggested that Spanish-speaking children ages one and older demonstrated lower preliminary levels of language development across time metrics, when compared to children whose primary language was English. Additionally, negative and statistically significant intercept beta coefficients indicated that children ages four and older whose primary language was Spanish demonstrated lower preliminary academic abilities across time metrics when compared to children whose primary language was English. Collectively, these results suggest that Spanish-speaking children a) tend to demonstrate lower patterns of language development across age bands, and b) may experience additional adverse effects on academic performance as academic rigor increases in pre-k and kindergarten.

Finally, despite the fact that several researchers have found that children whose primary language is Spanish typically demonstrate stronger developmental in early childhood (Hujar et al., 2021; Roberts & Bryant, 2011), results which primarily included nonsignificant slope beta coefficients suggested that there were not any consistent patterns of differences in growth rates between children whose primary language was Spanish and children whose primary language was English across age bands and domains of development. Together, these results suggest gaps are not systematically widening or narrowing between English-speaking and Spanish-speaking children from ages birth through six.

Children who Speak Other Languages.

Researchers have suggested that primary language is significantly related to children’s academic and developmental growth (Hujar et al., 2021; Roberts & Bryant, 2011), yet growth trends differ between children who belong to different primary language groups (Roberts & Bryant, 2011). Given within-group and between-cohort heterogeneity there were no a priori hypotheses regarding patterns of preliminary performance or growth for children whose primary language was categorized as *other.* Results including negative and statistically significant intercept beta coefficients for two- and three-year-old and kindergarten children suggested that children at these age levels demonstrated significantly lower preliminary developmental levels across most domains, when compared to their English-speaking peers. However, positive and statistically significant slope beta coefficients for two- to three-year-old, preschool, and kindergarten suggested that children at these age levels demonstrated significantly stronger growth for many domains of development than their English-speaking peers.

Black Non-Hispanic Children.

Research generally suggests that Black children demonstrate lower abilities early on (Kuhfeld et al., 2020; Young et al., 2017) that persist over time (Young et al., 2017). Therefore, it was hypothesized that Black Non-Hispanic children would demonstrate patterns of lower preliminary performance and growth over time. Results including negative and statistically significant intercept beta coefficients for children ages one and older generally suggested that Black Non-Hispanic children demonstrated significantly lower preliminary developmental levels across most domains of development than their White Non-Hispanic peers. However, results including non-significant, positive, and negative slope beta coefficients suggested that there was not a clear pattern of differences in growth rates between Black Non-Hispanic children and White Non-Hispanic children across age bands, domains, and time metrics.

Hispanic Children.

Research generally suggests that Hispanic children demonstrate lower abilities in early childhood (Reardon & Portilla, 2016; Reardon & Galindo, 2009) and stronger rates of growth prior to first or second grade (Hujar et al., 2021; Roberts & Bryant, 2011). Therefore, it was hypothesized that Hispanic children would demonstrate lower preliminary developmental levels and stronger rates of growth over time. While results, including negative intercept beta coefficients suggested that Hispanic children ages two and older demonstrated significantly lower preliminary developmental levels across most domains of development when compared to their White Non-Hispanic peers, nonsignificant slope beta coefficients suggested that Hispanic children did not experience significantly different growth rates when compared to their White Non-Hispanic peers. Collectively, these results suggest that while Hispanic children tend to demonstrate lower preliminary levels of development than their White Non-Hispanic peers, gaps are not systematically narrowing or widening from ages birth through age six.

Implications

Over time many educational researchers have examined differences in growth trends between subgroups of children, such as boys and girls (Hujar et al., 2021; Voyer et al., 2014) children of color and White children (Kuhfeld et al., 2020 Reardon & Portilla, 2016; Reardon & Galindo, 2009; Young et al., 2017) and children from diverse socioeconomic backgrounds (Hujar et al., 2021; Roberts & Bryant, 2011). By exploring differences in patterns of performance and growth over time, researchers have effectively established evidence of systematic differences in performance and growth between subgroups of children and influenced state- and federal-level policies. For example, President Lyndon B. Johnson established Head Start in 1964 to close opportunity gaps between children of color and White children and children living in poverty and more affluent children (Hudson, 2015). However, once policies are enacted, researchers must continually examine assessment data to understand whether current programs and initiatives are positively impacting children and closing opportunity and achievement gaps (Kuhfeld et al., 2020).

Results for the current study including, negative intercept beta coefficients for Black children, Hispanic children, and children whose primary language was not English suggested that these subgroups of children demonstrated lower preliminary levels of development than their White Non-Hispanic and English-speaking peers. These results provided support for study hypotheses and agreed with the broader literature on developmental differences for subgroups of children (Kuhfeld et al., 2020; Reardon & Portilla, 2016; Reardon & Galindo, 2009; Young et al., 2017). Additionally, positive intercept beta coefficients for girls suggested that girls demonstrated higher preliminary levels of development when compared to boys. Again, these results provided support for study hypotheses and agreed with the broader literature on developmental differences between boys and girls in early childhood and beyond (Kuhfeld et al., 2020; Voyer & Voyer, 2014). Together, these results suggest that subgroups of children, including children of color and White children, children from diverse linguistic backgrounds, and boys and girls continue to demonstrate significantly different patterns of development throughout early childhood.

Furthermore, while intercept beta coefficients generally revealed patterns of preliminary performance that were consistent with the broader literature, patterns in slope coefficients were less consistent across age bands, domains, and time metrics. For example, while it was hypothesized that girls would demonstrate stronger patterns of growth over time across age bands, results indicated that only kindergarten girls experienced stronger growth rates, when compared to boys. Furthermore, while it was hypothesized that children whose primary language was Spanish would demonstrate stronger patterns of growth over time across age bands, results suggested that there were no systematic patterns of differences across domains. While definitive conclusions cannot be made, there are several plausible explanations for the observed results.

First, while the current study sought to establish a racially and ethnically representative sample of children, the number of children who spoke languages other than English varied by cohort. For example, while 726 kindergarteners spoke *other* languages, only 250 pre-k children spoke *other* languages. Given that some age bands had larger samples of children with particular characteristics, statistical significance was easier to obtain for some age bands than others. Second, given that data are from age-separated cohorts of children, it's possible that there were true differences in patterns of growth over time between cohorts of children. Third, large proportions of residual variance for one- to two-year-old, two- to three-year-old, and preschool models suggested that there were more significant rater effects for some age bands, making it more challenging to detect true differences in growth rates. And finally, the absence of consistent patterns of differences in growth over time for most subgroups of children prior to kindergarten or first grade could suggest that while gaps exist, as evidenced by statistically significant intercept beta coefficients, gaps are not systematically narrowing or widening between subgroups of children from birth through age six. For example, even though Spanish-speaking children demonstrate patterns of lower preliminary language development from age one through kindergarten, nonsignificant slope coefficients for most age bands suggest that Spanish-speaking children are still acquiring language skills at the same rate as their English-speaking peers.

Finally, while many researchers have found that subgroups of children demonstrate significantly different patterns of performance at school entry across domains of learning and development (Kuhfeld et al., 2020; Reardon & Galindo, 2009; Reardon & Portilla, 2016), less is known about when differences arise. For example, researchers have commonly observed that Black and Hispanic kindergarten children demonstrate lower abilities at kindergarten entry than their White peers (Kuhfeld et al., 2020; Reardon & Portilla, 2016), yet less has been written about when the gaps in performance begin to emerge. Results from the present study suggest that subgroups of children begin to demonstrate different patterns of performance around one- or two-years of age. For example, children whose primary language is Spanish and children who identify as Black Non-Hispanic demonstrate lower preliminary developmental levels around age one, when compared to White Non-Hispanic children and children whose primary language is *other* and children who identify as Hispanic demonstrate lower preliminary developmental levels around age two. While girls tend to demonstrate stronger patterns of performance, when compared to boys, these gaps in performance also emerge around age two. Together these results suggest that gaps between subgroups of children begin to emerge by one to two years of age and children may need access to high-quality care prior to age one or two to mitigate opportunity gaps and observable developmental differences at school entry.

Recommendations

Subgroups of children including, girls, Black children, Hispanic children, and children whose primary language was not English, demonstrated significantly different patterns of preliminary performance across domains of learning and development beginning around age one or two, when compared to the reference group. Given these results and extant research which suggests that children as young as two years old are profoundly influenced by caregiver interactions (Gialamas et al., 2013) and the learning environment (Yoshikawa et al., 2013), both state and federal educational agencies should seek to provide more comprehensive high-quality early childhood education for children prior to one- or two-years of age. By increasing access to high-quality care for subgroups of children who commonly demonstrate lower patterns of performance, including boys, children who are Black or Hispanic, and children who speak languages other than English at home, we may be able to reduce the differences in patterns of performance that are evident at school entry (Kuhfeld et al., 2020; Reardon & Portilla, 2016; Reardon & Galindo, 2009) and persist through twelfth grade and beyond (Shanley et al., 2016; Mok et al., 2015; Voyer & Voyer, 2014, Young et al., 2017).

Next, although patterns of differences in preliminary performance were generally in the hypothesized direction, patterns were somewhat inconsistent across age bands, domains, and time metrics for some subgroups of children. Observed differences between age-separated cohorts could be attributed to a number of factors, including a) true differences in patterns of performance between age-separated cohorts of children, b) different sample sizes for some characteristics (e.g., Spanish-speaking children), c) more significant rater effects for some age bands or d) confounding variables (e.g., disability status or socioeconomic status). To strengthen inferences about the relationships between subgroup membership and patterns of performance and growth over time, future research should seek to a) replicate the present study with multi-year longitudinal data for one cohort of children to eliminate competing explanations for different results between age bands, b) consider using assessment records from children whose teachers are interrater reliability certified to eliminate some degree of construct irrelevant variance, and c) model additional influential child-level characteristics, including socioeconomic status and disability status. By replicating the present study with more reliable assessment data and additional child-level characteristics, researchers may be able to detect true differences in patterns of performance and growth more easily.

Finally, while I do not recommend using subgroup-specific slope and intercept beta coefficients to construct subgroup-specific normative scores due to the limitations of the current study discussed above, in the presence of more stable estimates, future researchers may construct subgroup-specific norms to help educational stakeholders understand typical patterns of performance given influential child-level characteristics. While subgroup-specific normative scores would not necessarily aid teachers in understanding relative performance and making data-driven decisions, they may help to highlight gaps in preliminary performance and growth over time between subgroups of children and substantiate the need for additional services and programs for young children who tend to experience different patterns of performance, when compared to age and grade level peers.

Summary

The primary purpose of the current study was to use HLMs to establish time-continuous national normative scores for the most widely used authentic formative assessment in early childhood, *GOLD*®. Secondary study purposes included, a) identifying the most effective time metric for modeling growth across domains of learning and development for children ages birth through kindergarten, b) making inferences about the nature of the growth process (e.g., shape of the growth trajectory), and c) exploring whether or not and to what extent child-level characteristics were associated with different growth trajectories.

Results from the present study seek to make several significant contributions to both research and practice. First, results corresponding to the first research question, including several sources of validity evidence suggested that HLM-estimated normative ability and growth scores are reasonably valid and reliable. This finding is important for two reasons, including, a) valid and reliable national normative scores are prerequisite to making appropriate, useful, and meaningful relative score interpretations, and b) results provide additional support for the use of MLMs to estimate normative ability and growth scores (Pan & Goldstein, 1997; Thum & Kuhfeld, 2020). Next, results corresponding to the second research question generally suggest that age is the most effective predictor of developmental growth for children in one- to two-year-old classrooms and instructional exposure is the most effective predictor of developmental growth for children in preschool, pre-k, and kindergarten. Although future research should seek to model instructional exposure with greater precision to strengthen analyses and confirm results, together these results generally agree with the broader literature and provide further evidence of the importance of high-quality instruction for children ages three and older. Results corresponding to the third research question generally suggest that developmental growth, as measured by *GOLD*®, slows across the first three age bands and accelerates across the next three age bands. Collectively, these results suggested that developmental growth, as measured by *GOLD*®, is non-linear and quadratic models may demonstrate good fit to multi-year longitudinal data for children ages birth through three and children ages three through six. And finally, results corresponding to the final research question, suggest that subgroups of children, including Black children, Hispanic children, children whose primary language is not English, and boys demonstrate significantly lower patterns of preliminary performance beginning around age one or two. Together, these results suggest that differences in patterns of performance begin as early as one year of age, and young children may need access to high-quality care from infancy to mitigate gaps between subgroups of children that are commonly observed at kindergarten entry (Kuhfeld et al., 2020; Reardon & Portilla, 2016; Reardon & Galindo, 2009) and persist through twelfth grade and beyond (Shanley et al., 2016; Mok et al., 2015; Voyer & Voyer, 2014, Young et al., 2017).

REFERENCES

Akaeze, H. O., Lawrence, F. R., & Wu, J. H. C. (2022). Resolving dimensionality in a child assessment tool: An application of the multilevel bifactor model. *Educational and Psychological Measurement, 00*(0), 1-23. https://doi: 10.1177/00131644221082688

The Annie E. Casey Foundation. (2023, March 27). Child population by race and ethnicity in the United States. *Kids Count Data Center.* https://datacenter.kidscount.org/data/tables/103-child-population-by-race-andethnicity#detailed/1/any/false/2048,574,1729,37,871,870,573,869,36,868/68,69,67,12,70,66,71,72/423,424

American Educational Research Association, American Psychological Association, & National Council of Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: AERA.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores.* Educational Testing Service.

Bagnato, S. J., Goins, D. D., Pretti-Frontczak, K., & Neisworth, J. T. (2014). Authentic assessment as "Best practice" for early childhood intervention. *Topics in Early Childhood Special Education*, *34*(2), 116–127. https://doi.org/10.1177/0271121414523652

Bagnato, S. J., & Ho, H. Y. (2006). High stakes testing with preschool children: Violation of professional standards for evidence-based practice in early childhood intervention. *KEDI International Journal of Educational Policy*, *3*(1), 22–43.

Bredekamp, S., & Copple, C. (2008). *Developmentally appropriate practice in early childhood programs* (3rd ed.). National Association for the Education of Young Children

Blewitt, C., O'Connor, A., Morris, H., Nolan, A., Mousa, A., Green, R., Ifanti, A., Jackson, K., & Skouteris, H. (2021). "It's embedded in what we do for every child": A qualitative exploration of early childhood educators' perspectives on supporting children's social and emotional learning. *International Journal of Environmental Research and Public Health*, *18*(4), 1530–1546. https://doi.org/10.3390/ijerph18041530

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach.* John Wiley & Sons.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*(1), 147-158. https://doi.org/10.1037/0033-2909.101.1.147

Bryk, A. S., & Weisberg, H. I. (1976). Value-added analysis: A dynamic approach to the estimation of treatment effects. *Journal of Educational Statistics. 1*(2), 127-155. https://doi.org/10.2307/1164980

Burts, D. B., Berke, K., Heroman, C., Baker, H. Bickart, T., Tabors, P., & Sanders, S. (2016). *GOLD® objectives for learning & development: Birth through third grade.* Teaching Strategies®.

Center on the Developing Child at Harvard University (2016). *From best practices to breakthrough impacts: A science-based approach to building a more promising future for young children and families.* http://www.developingchild.harvard.edu

Chetty, R., Friedman, J. N., Hilger, N., Saez, E., Schanzenbach, D. W., & Yagan, D. (2011). How does your kindergarten classroom affect your earnings? Evidence from Project Star. *The Quarterly Journal of Economics*, *126*(4), 1593–1660. https://doi.org/10.1093/qje/qjr041

Cole, D. A. & Preacher, K. J. (2013). Manifest Variable Path Analysis: Potential Serious and Misleading Consequences Due to Uncorrected Measurement Error. *Psychological*

Methods, 19(2), 300-315. https://doi.10.1037/a0033805

Corcoran, R. P., Cheung, A. C., Kim, E., & Xie, C. (2018). Effective universal school-based social and emotional learning programs for improving academic achievement: A systematic review and meta-analysis of 50 years of research. *Educational Research Review*, *25*, 56–72. https://doi.org/10.1016/j.edurev.2017.12.001

Curran, P. J. (2003) Have multilevel models been structural equation models all along?. *Multivariate Behavioral Research, 38*(4), 529-569. https://doi.org/10.1207/s15327906mbr3804_5

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development. 11*(2), 121-136. https://doi.org/10.1080/15248371003699969

DeSilver, D. (2019). *'Back to school' means anytime from late July to after Labor Day, depending on where in the U.S. you live.* Pew Research Center. https://www.pewresearch.org/fact-tank/2019/08/14/back-to-school-dates-u-s/

Diakow, R. (2018). Growth curve modeling. In B. B. Frey (Ed.), *The Sage encyclopedia of educational research, measurement, and evaluation* (no page number). SAGE Publications. https://doi.org/10.4135/9781506326139

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Routledge. https://doi.org/https://10.4324/9780203874776.ch1

Ebel, R. L. (1962) Content standard test scores. *Educational and Psychological Measurement, 22*(1), 15-25. https://doi.org/10.1177/001316446202200103

Education Department (2016, August 8). Applications for new awards; Enhanced assessment instruments grant program-enhanced assessment instruments. *The Daily Journal of the U.S. Government.* https://www.federalregister.gov/documents/2016/08/08/2016

Executive Office of the President (2002). *Good Start, Grow Smart: The Bush Administration's early childhood initiative.* ERIC.

Fantuzzo, J., Hightower, D., Grim, S., & Montes, G. (2002). Generalization of the Child Observation Record: A validity study for diverse samples of urban, low-income preschool children. *Early Childhood Research Quarterly, 17*(1), 106-125. https://doi.org/10.1016/s0885-2006(02)00131-X

Ferrer, E., Hamagami, F., & McArdle, J. J. (2009). Modeling latent growth curves with incomplete data using different types of structural equation modeling and multilevel software. *Structural Equation Modeling A Multidisciplinary Journal, 11*(3), 452-483. https://doi.org/10.1207/s15328007sem1103_8

Gialamas, A., Mittinty, M. N., Sawyer, M. G., Zubrick, S. R., & Lynch, J. (2013). Childcare quality and children's cognitive and socio-emotional development: An Australian Longitudinal Study. *Early Child Development and Care*, *184*(7), 977–997. https://doi.org/10.1080/03004430.2013.847835

Goldstein, J., & Flake, J. K. (2015). Towards a framework for the validation of early childhood assessment systems. *Educational Assessment, Evaluation, and Accountability, 28*(3), 273-293. https://10.1007/s11092-015-9231-8

Hamilton, J. (2022). *Scans reveal the brain's early growth, late decline and surprising variability.* https://www.npr.org/sections/health-shots/2022/04/07/1091309647/brain-development-disorder

Hanover Research (2013). *Kindergarten entry assessments: Practices and policies.*

https://www.hanoverresearch.com/media/Kindergarten-Entry-Assessments-Practices-and-Policies.pdf

Harkness, S., Super, C. M., Johnston Mavridis, C., Barry, O., & Zeitlin, M. (2013). Culture and early childhood development: Implications for policy and programs. In P. Rebello Britto, P. L. Engle, & C. M. Super (Eds.), *Handbook of early childhood development research and its impact on global policy* (1st ed., pp. 142–160). Oxford University Press.

Hartwig, E. A. (2016, December 7). *Authentic assessment: A critical tool for early childhood educators*. New York Early Childhood Professional Development Institute. https://earlychildhoodny.org/blog/authentic-assessment-a-critical-tool-for-early-childhood-educators.

H.R. 2210: The school readiness act of 2003: Hearing before the Subcommittee on Education Reform of the Committee on Education and the workforce, House of Representatives, One Hundred eighth Congress, First Session, hearing held in Washington, DC, June 3, 2003 (2003). bill.

Hudson, D. (2015). This day in history: The creation of Head Start. *The White House.* https://pbamawhitehouse.archives.gov

Hujar, J., Luce, H. E., & Lambert, R. G. (2021). Relationship between *GOLD*® interrater reliability certification status and teacher ratings of child developmental progress using the *GOLD*® assessment [Unpublished manuscript]. Department of Educational Leadership, The University of North Carolina at Charlotte.

Kim, D. H., Lambert, R. G., & Burts, D. C. (2013). Evidence of the validity of *Teaching*

Strategies GOLD® assessment tool for English language learners and children with disabilities. *Early Education and Development*, *24*(4), 574–595. http://doi.org/10.1080/10409289.2012.701500

Klein, A. (2016, March 31). *The Every Student Succeeds Act: An ESSA overview.* Education Week. https://www.edweek.org/policy-politics/the-every-student-succeeds-act-an-essa-overview/2016/03

Klein, A. (2015, April 10). *No Child Left Behind: An overview*. Education Week. https://www.edweek.org/policy-politics/no-child-left-behind-an-overview/2015/04

Kohli, N., Sullivan, A. L., Sadeh, S., & Zopluoglu, C. (2015). Longitudinal mathematics development of students with learning disabilities and students without disabilities: A comparison of linear, quadratic, and piecewise linear mixes effects models. *Journal of School Psychology, 53*(2), 105-120. https://doi.org/10.1016/j.jsp.2014.12.002

Kuhfeld, M., Soland, J., Pitts, C., & Burchinal, M., (2020). Trends in children's academic skills at school entry from 2010 to 2017. *Educational Researcher, 49*(6), 403-414. https://doi.org/10.3102/0013189X20931078

Lambert, R. G. (2022, December). *Technical manual for the Teaching Strategies GOLD® assessment: Birth through third grade edition (3rd edition)*. Center for Educational Measurement and Evaluation, University of North Carolina Charlotte.

Lambert, R. G. (2020, October). *Technical manual for the Teaching Strategies GOLD® Assessment (2nd edition) birth through third grade*. Center for Educational Measurement and Evaluation, University of North Carolina Charlotte. https://teachingstrategies.com/wpcontent/uploads/2020/09/2020-Tech-Report_GOLD_B3_V4.pdf?msclkid=7fd30ba0ab8911ecb3c764a7ab95c22c

Lambert, R. G. (2017). *Technical manual for the Teaching Strategies GOLD® assessment system: Birth through third grade edition.* Center for Educational Measurement and Evaluation, University of North Carolina Charlotte.

Lambert, R. G., Kim, D. H., & Burts, D. C. (2015). The measurement properties of the *Teaching Strategies GOLD®* assessment system. *Early Childhood Research Quarterly*, *33*(1), 49–63. https://doi.org/10.1016/j.ecresq.2015.05.004

Lee, J. (2010). Tripartite growth trajectories of reading and mathematics achievement: Tracking national academic progress at primary, middle, and high school levels. *American Educational Research Journal. 47*(4), 800-832. https://doi.org/10.3102/0002831210365009

Lee, J. H. & Huber, J. C. (2021). Evaluation of multiple imputation with large proportions of missing data: How much is too much? *Iranian Journal of Public Health, 50*(7), 1372-1380. https://doi:10.18502/ijph.v50i7.6626

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: Compatibility and complementary. *Assessment & Evaluation in Higher Education, 41*(3), 450-465. https://doi.org/10.1080/02602938.2015.1022126.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement*, *16*(4), 421-437. https://doi.org/10.1177/001316445601600401

Marrs, A. (2020, November). *Whole child assessment*. Kaymbu360.

Masonic Institute for the Developing Brain. (2021). *First 1,000 days – A critical time for children's brain development.* https://mhealthfairview.org/blog/first-1000-days-critical-time-for-brain development?utm_source=twitter&utm_medium=social_organic&utm_campaign=ummch_comms_midb_childhood_stress_trauma

McCoach, D. B., Newton, S. D., & Gambino, A. J. (2022). Multilevel Model Selection: Balancing Model Fit and Adequacy. In M.S. Khine (ed.), *Methodology for Multilevel Modeling in Educational Research* (pp. 29-48). Springer. https://doi.org/10.1007/978-981-16-9142-3_3

McNeish, D. & Matta, T. (2017). Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Behavioral Research. 50,* 1398-1414. https://doi.10.3758/s13428-017-0976-5

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5*(1)*,* 23-43. https://doi:10.1037//1052-989X.5.I.2.3

Miller, P., Podvysotska, L. B., & Votruba-Drzal, E. (2021). Wealth inequality and child development: Implications for policy and practice. *The Russell Sage Foundation Journal of the Social Sciences, 7*(3), 154-174. https://doi.org/10.7758/rsf.2021.7.3.07.

Mok, M. M., McInerney, D. M., Zhu, J., & Or, A. (2015). Growth trajectories of mathematics achievement: Longitudinal tracking of student academic progress. *British Journal of Educational Psychology*, *85*(2), 154–171. https://doi.org/10.1111/bjep.12060

Muthen, B. O., & Curran, P. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analyses and power estimation. *Psychological Methods, 2*(4)*,* 371-402.

National Association for the Education of Young Children (2020). *Developmentally appropriate practice position statement.* https://www.naeyc.org/resources/position-statements/dap/contents

National Association for the Education of Young Children (2018). *Staff-to-Child Ratio and Class*

Size. https://www.naeya.org/sites/defaults/files/globally-shared/downloads/PDFs/accreditation/early-learning/staff_child_ratio_0.pdf

National Research Council (2008). *Early childhood assessment: What, why, and how.* https://www.acf.hhs.gov/sites/default/files/documents/opre/early_child_brief.pdf

O'Connell, A. A., & McCoach, D. B. (2008). *Multilevel modeling of educational data.* Information Age Publishing Inc.

O'Connor, G. (2019). *A kindergarten age guide for parents: When kids should start school.* https://www.parents.com/kids/education/kindergarten/kindergarten-age-guide-for-parents/

The Office of Early Childhood Development. (2019, October 18). *Race to the Top – Early Learning Challenge.* https://www.acf.hhs.gov/ecd/early-learning/race-top.

Pace, A., Alper, R., Burchinal, M. R., Golinkoff, R. M., & Hirsh-Pasek, K. (2019). Measuring success: Within and cross-domain predictors of academic and social trajectories in elementary school. *Early Childhood Research Quarterly, 46*(1), 112-125. https://doi.org/10.1016/j.ecresq.2018.04.001

Palmeri, M. (n.d.). *Testing the assumptions of multilevel models.* A Language not a Letter: Learning Statistics in R. Retrieved July 6, 2023, from https://ademos.people.uic.edu/Chapter18.html

Pan, H. & Goldstein, H. (1997). Multilevel models for longitudinal growth norms. *Statistics in Medicine, 16,* 2665-2678.

Piaget, J. (1971) The theory of stages in cognitive development. In D. R. Green, M. P. Ford, & G. B., Flamer (Eds.), *Measurement and Piaget* (pp. 1-11). McGraw Hill.

Pirralha, A. (2020). *Testing for measurement invariance with many groups.* Bookdown.

https://bookdown.org/andrepirralha/bookdown-demo/

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* (2nd ed.). Sage.

Reardon, S. F., & Galindo, C. (2009). The Hispanic-White achievement gap in math and reading in the elementary grades. *The American Research Journal, 46*(3), 853-891. https://doi.org/10.3102/0002831209333184

Reardon, S. F., & Portilla, X. A. (2016). Recent trends in income, racial, and ethnic school readiness gaps at kindergarten entry. *AERA Open, 2*(3), 1-18. https://doi:10.1177/2332858416657343.

Rebello-Britto, P. R., Engle, P. L., & Super, C. M. (2013). Early childhood development: Translating research to global policy. In P. Rebello Britto, P. L. Engle, & C. M. Super (Eds.), *Handbook of early childhood development research and its impact on global policy* (1st ed., pp. 3–23). Oxford University Press.

Rhoades, B. L., Warren, H. K., Domitrovich, C. E., & Greenberg, M. T. (2011). Examining the link between preschool social-emotional competence and first grade academic achievement: The role of attention skills. *Early Childhood Research Quarterly, 26*(2), 182-191. https://doi:10.1016/j.ecresq.2010.07.003

Roberts, G., & Bryant, D. (2011). Early mathematics achievement trajectories: English-language learner and native English-speaker estimates, using the Early Childhood Longitudinal Survey. *Developmental Psychology, 47*(4), 916-930. https://doi:10.1037/a0023865.

Rogosa, D. R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*(3)*,* 726-748.

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Education Measurement, 20*(4)*,* 335-343.

Shanley, L. (2016). Evaluating longitudinal mathematics achievement growth. *Educational Researcher*, *45*(6), 347–357. https://doi.org/10.3102/0013189x16662461

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

Shin, T., Davison, M. L., Long, J. D., Chan, C.-K., & Heistad, D. (2013). Exploring gains in reading and mathematics achievement among regular and exceptional students using growth curve modeling. *Learning and Individual Differences*, *23*, 92–100. https://doi.org/10.1016/j.lindif.2012.10.002

Stipek, D. (2006). No child left behind comes to preschool. *The Elementary School Journal*, *106*(5), 455–466. https://doi.org/10.1086/505440

Thum, Y. M., & Kuhfeld, M. (2020). *NWEA 2020 MAP growth achievement status and growth norms for students and schools.* NWEA. https://teach.mapnwea.org/impl/normsResearchStudy.pdf

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140*(4), 1174-1204. https://dx.doi.org/10.1037/a0036620.

Wakabayashi, T., Claxton, J., & Smith, E. V. (2019). Validation of a revised observation-based assessment tool for children birth through kindergarten: The COR Advantage. *Journal of Psychoeducational Assessment, 37*(1), 69-90. https://doi.10.1177/0734282917732491

Young, J. L., Young, J. R., Ford, D. Y. (2017). Standing in the gaps: Examining the effects of early gifted education on Black girl achievement in STEM. *Journal of Advanced*

Academics, 28(4), 290-312. https://doi.org/10.1177/1932202X17730549

Yoshikawa, H., Weiland, C., Brooks-Gunn, J., Burchinal, M. R., Espinosa, L. M., Gormley, W. T., Ludwig, J., Magnuson, K. A., Phillips, D., & Zaslow, M. J. (2013). *Investing in our future: The evidence base on preschool education.* Society for Research on Child Development & Foundation for Child Development. https://files.eric.ed.gov/fulltext/ED579818.pdf

APPENDIX: DEMOGRAPHIC CHARACTERISTICS OF THE FULL SAMPLE

A1

Age in Months on September 1, 2021

Age/Grade Band	Age in Months				
	n	*M*	*SD*	Minimum	Maximum
Birth to One-Year Olds	3000	4.38	2.16	0	8
One- to Two-Year Olds	3000	15.56	2.28	12	19
Two- to Three Year Olds	3000	27.34	2.30	24	31
Preschool	3000	41.12	3.10	36	46
Pre-K	3000	52.90	3.15	48	58
Kindergarten	3000	64.71	3.23	60	70

A2

Demographic Characteristics of the Full Sample

Characteristic		*n*	%
Gender	Female	8655	48.12
	Male	9330	51.88
Race/Ethnicity	White	8983	49.91
	Black	2484	13.80
	Asian	890	4.94
	Native American, Pacific Islander, Alaska Native & Hawaiian Native	187	1.04
	Multiple Races	832	4.62
	Hispanic	4624	25.69
Primary Language	English	13429	74.61
	Spanish	2454	13.63
	Other	2117	11.76
ELL Status	Identified as an ELL	1635	9.08
Disability Status	Has IFSP or IEP	1226	6.81
Poverty Status	Receives FRL	4993	27.74