African American English and early literacy: A comparison of approaches to quantifying nonmainstream dialect use

Zachary Maher[1,2*]

Michelle Erskine[2]

Arynn Byrd[2]

Jeffrey Harring[3]

Jan Edwards[2,4]

[1]Program in Neuroscience and Cognitive Science

[2]Department of Hearing and Speech Sciences

[3]Department of Human Development and Quantitative Methodology

[4]Maryland Language Science Center

University of Maryland, College Park; College Park, MD

*Corresponding Author

zach@umd.edu

0121 Taliaferro Hall, Chapel Drive, College Park, MD 20724

(301)-405-9789

Conflict of Interest

The authors claim no conflicts of interest.

Funding Statement

This research was supported in part by IES grant #R305A17013 to Jan Edwards and NSF grant #1449815 to Colin Phillips.

Abstract

Purpose: Many studies have found a correlation between overall usage rates of nonmainstream forms and reading scores, but less is known about which dialect differences are most predictive. Here, we consider different methods of characterizing AAE use from existing assessments and examine which methods best predict literacy achievement.

Method: Kindergarten and first grade students who speak AAE received two assessments of dialect use and two assessments of decoding at the beginning and end of the school year. Item-level analyses of the dialect-use assessments were used to compute measures of dialect usage: (1) an overall feature rate measure based on the DELV-ST, (2) a subscore analysis of the DELV-ST based on items that pattern together; (3) an alternative assessment where children repeat and translate sentences; and (4) "repertoire" measures based on a categorical distinction of whether a child used a particular feature of MAE.

Results: Models using feature rate measures provided better data-model fit than those with repertoire measures, and baseline performance on a sentence repetition task was a positive predictor of reading score at the end of the school year. For phonological subscores, change from the beginning to end of the school year predicted reading at the end of the school year, while baseline scores were most predictive for grammatical subscores.

Conclusions: The addition of a sentence imitation task is useful for understanding a child's dialect and anticipating potential areas for support in early literacy. We observed some support for the idea that morphological dialect differences (i.e. irregular verb morphology) have a particularly close tie to later literacy, but future work will be necessary to confirm this finding.

Introduction

Since the earliest sociolinguistic work on nonmainstream varieties of American English, there has been considerable interest in potential educational implications of dialect differences. Many correlational relationships have been observed between a variety of measures of dialect difference and a variety of literacy outcomes, and multiple curricula have been proposed to support emergent readers who speak nonmainstream dialects.

All of this work requires that researchers (1) have a framework for understanding what dialect variation is and (2) operationalize this understanding with one or more measures of participants' dialect use. Both of these steps are fraught with challenges and require simplification that will fail to fully capture individuals' experience with linguistic variation. Such simplification can affect the inferences we draw about the relationship between dialect differences and early literacy and, in turn, affect the strategies we use to support emergent readers.

Perhaps the most common measurement of children's dialect use is Part 1 of the Diagnostic Evaluation of Language Variation-Screening Test (DELV-ST; Seymour, Roeper, de Villiers, & de Villiers, 2003). This measure was designed to determine whether a child speaks a nonmainstream dialect of American English (NMAE), so it focuses on the most reliably produced nonmainstream features (primarily production of dental fricatives and subject-verb agreement patterns). Because of this, the DELV-ST is not ideal for capturing children's knowledge of Mainstream American English (MAE). This paper uses the DELV-ST along with a different test, the Dialect Assessment Battery (DAB; Craig, 2014), which is explicitly designed to see whether children can produce MAE-compatible features and if they can translate nonmainstream dialect features into MAE.

We will begin with an overview of different approaches to characterizing dialect differences, with the goal of adapting our measures to capture some of the insights from these approaches. Next, we will review research on the relationship between dialect differences and early literacy, which motivates the present study, with a focus on studies using the DELV-ST. We will then present results from an ongoing study of speakers of African American English (AAE) in kindergarten and first grade, comparing different approaches to measuring their linguistic system and reporting the implications that these approaches have on predicting growth in decoding scores.

Background

Characterizing Variation

One common approach to quantifying AAE involves the use of lists of AAE features, creating a Dialect Density Measure (DDM) that corresponds to the rate of usage of such features. For example, Washington and Craig (2002) used a list of 26 features of AAE that differ from MAE, including zero copula (*They not finished eatin' yet*), multiple negation (*I don't remember nobody havin' no motorcycle*), and variation in subject-verb agreement (*I knew you was gonna say that*). They then calculated DDM as the number of tokens of any of these features divided by the number of words in a given language sample. Despite the widespread usage of this approach under the "DDM" terminology, we will use the more recent term Nonmainstream Form Density (NMFD), following others such as McDonald and Oetting (2019), to emphasize the fact that everyone speaks a dialect, and feature rate methods inherently involve a comparison to a perceived standard, the "mainstream dialect."

Variants of this approach have been used with tasks ranging from highly structured sentence-elicitation tasks (e.g., Charity et al., 2004), to open-ended narrative tasks (e.g., Renn &

Terry, 2009; Craig et al., 2014). Multiple approaches to feature sets have been shown to be highly correlated; for example, Renn and Terry (2009) found that a subset of just six features provided comparable results to a 40-feature list in detecting style shifts in AAE-speaking sixth graders, Oetting and McDonald (2002) found that type-based and token-based approaches correctly categorized child speakers of AAE and Southern White English, and Oetting and Pruitt (2005) found that streamlined feature lists can often be sufficient for characterizing participants' dialect. Multiple approaches have also been used for the denominator in these calculations, including number of opportunities to use a feature (in more structured tasks), number of words in the sample (e.g., Washington & Craig, 2002; Horton-Ikard & Weismer, 2005), and number of utterances in the sample (e.g., Craig & Washington, 2000). NMFD approaches have also been used to study changes in dialect use over time. For example, Terry and Connor (2012) found a decrease in NMFD between kindergarten and first grade, and Terry, Connor, Petscher, and Ross Conlin (2012) found a decrease in NMFD over the course of first grade that leveled off in second grade. These changes likely represent a combination of factors, including development within AAE toward a more adultlike grammar, developing knowledge of MAE, and changes in style-shifting (Beyer & Hudson Kam, 2012; Green, 2011).

While the feature rate approach is helpful for quantifying differences from MAE, it has many limitations, which are discussed at length by Green (2011, Ch. 2). She argues that feature lists are ill-equipped to characterize AAE as its own rule-governed system. This means, for example, that groups of features might share underlying patterns, and NMFD would not highlight this fact. Additionally, NMFD measures typically treat features that are consistent with MAE as use of MAE, while only features of AAE that differ from MAE are counted as AAE use. This might be appropriate for verbal *-s*, which is often considered to be absent from the AAE grammar (Newkirk-Turner & Green, 2016; for criticism of this view, see Baugh, 1990; Cleveland, & Oetting, 2013; Barriere et al., 2019). However, it is often the case that the "MAE" feature is also available within the grammar of AAE. For example, "zero copula" commonly appears on AAE feature lists, but it is variable, with overt copulas also being acceptable in AAE (e.g., Wyatt, 1996; Roy, Oetting, & Moland, 2013; Newkirk-Turner, Oetting, & Stockman, 2014).

The feature rate approach also fails to account for insights from third wave sociolinguistics (Eckert, 2008, 2012). Third wave sociolinguistics proceeds from the idea that speakers have a range of sociolinguistic variables that they can deploy in different social situations. These variables pattern together as styles, which give variables their social meanings, and individuals use styles to express their ideologies about membership in different groups. This approach was not necessarily developed with a focus on AAE, but it provides a framework to appreciate the more nuanced nature of dialect variation and has been applied to more recent work characterizing the language of African Americans (e.g., King, 2018). This is in contrast to early work in sociolinguistics, which often sought to characterize idealized vernacular forms, such as focusing on the idea of a pure speaker of AAE who does not "code-switch" into MAE (see discussion in Wolfram, 2007; King, 2020).

As Snell (2013) argues, the more recent work in sociolinguistics on styles allows us to use *repertoire* as an alternative framing for children's knowledge of variation. She points out that recent educational work tries to replace the "deficit narrative" (i.e. that nonmainstream features indicate poor language skills) with a "difference narrative," suggesting that nonmainstream varieties are distinct, rule-governed systems. However, both of these narratives make the assumption that discrete varieties of English exist, which is not borne out in the data. For example, she finds that 9- and 10-year-old children in northeast England mix regional and standard feature use within one discourse depending on how they are trying to socially position themselves relative to their interlocutors.

While the UK dialect context is different from that of the United States, neither could be characterized as strict diglossia, where there is clear separation between vernacular and standard dialects (e.g., Auer, 2005). Also, it is not necessarily the case that a child with good metalinguistic skills and knowledge of MAE will use MAE forms in the school setting; these children also have compelling reasons to assert a Black identity using their speech (Ogbu, 1999), and the use of different forms could be a means to assert social difference from the examiner, a process known as divergence in Communication Accommodation Theory (Giles & Ogay, 2007). Thus, it is possible that the mere presence of a particular MAE-compatible form (e.g., an overt copula) within a child's repertoire is more important than the rate at which the child favors the form over MAE-incompatible alternatives (e.g., a zero copula).

Dialect Differences and Literacy

Despite the criticisms of the NMFD approach, Van Hofwegen and Wolfram (2017) argue that aggregate NMFD values are still useful, particularly when trying to track individuals' changes in dialect usage over time and in large-scale, multidisciplinary studies in general. This is a likely source of their popularity in research on the relationship between dialect differences and literacy. Over the past two decades, a large body of work has developed to address the extent to which *dialect mismatch*—the presence of linguistic differences between nonmainstream dialects (e.g., AAE) and mainstream dialects (e.g., MAE)—impacts children's literacy achievement (e.g., Connor & Craig, 2006; Labov, 1995; Terry & Scarborough, 2011). The influence of dialect mismatch on literacy achievement spans various subcomponents of reading, including decoding and reading comprehension, though the majority of this research has focused on decoding. Studies vary in their commitments to models of reading, but we will assume the Simple View of Reading, which posits that reading is a product of decoding and linguistic comprehension (e.g., Gough & Tunmer, 1986; Hoover & Gough, 1990).

Research on language variation emerging from fields such as speech-language pathology, education, linguistics, and sociolinguistics posits an inverse relationship between reading achievement and the use of AAE. For example, Charity, Scarborough, and Griffin (2004) examined the relationship between children's facility with MAE via a sentence repetition task that was designed to elicit features of AAE, and reading performance using a standardized assessment of decoding (Woodcock Reading Mastery Test-Revised, Word Attack Subtest). They calculated two NMFD scores corresponding to phonological and morphological features of AAE, and they observed an inverse relationship between reading performance and the use of nonmainstream dialect features for both the phonological and morphological measures.

Shade (2012) also used separate NMFD measures for phonological and morphological features. She found that both measures were negatively correlated with decoding, but only phonological NMFD was predictive of sight word reading. Research on the relationship between dialect differences and literacy has generally not looked at finer-grained differences within the broader categories of phonological and morphological variation, though multiple studies report usage rates by feature (e.g., Washington & Craig, 2002; Craig, Thompson, Washington, & Potter, 2003; Oetting & McDonald, 2002). For example, Oetting and McDonald (2002) found that 100% of African American children used zero-marking on regular present tense verbs with third person singular subjects, but only 70% used zero marking for irregular verbs. Such separation of regular and irregular forms is also a longstanding finding in the acquisition literature (e.g., Brown, 1973).

Other studies have used more traditional assessment methods to examine the relationship between nonmainstream language variation and reading. Champion, Rosa-Lugo, Rivers, and McCabe (2010) used the DELV-ST to identify speakers of nonmainstream English varieties and to evaluate how performance on this screener related to a test of oral reading, the Gray Oral Reading Test-Fourth Edition (GORT-4). Children who produced a greater number of nonmainstream features had lower scores on the GORT-4. Others, such as Terry and Connor (2012), found that a change in performance on the DELV-ST across two time points was predictive of decoding skills. Children who decreased their use of nonmainstream features between kindergarten and first grade had higher reading scores. This finding also highlights the relevance of the time course of the relationship between NMAE use and changes in reading.

Collectively, this scholarship suggests that children who demonstrate higher nonmainstream dialect density and less facility with varying their use of MAE in different contexts (i.e., dialect-shifting) exhibit poorer literacy outcomes. This relationship remains true for studies examining emergent literacy skills, such as decoding, and later literacy skills, such as reading comprehension (Terry et al., 2016). Moreover, the established effects of dialect mismatch on reading are above and beyond socioeconomic differences and race, factors that were previously shown to obscure the relationship between dialect mismatch and reading achievement (Bühler, von Oertzen, McBride, Stoll, & Maurer, 2018).

Questions

Here, we contrast different scoring approaches to assessments that target NMFD. We asked the following research questions:

1. Do subsets of DELV-ST items pattern together in a way that corresponds to different components of the AAE system?
2. Does nonmainstream dialect usage at the beginning of the school year, or change in nonmainstream dialect usage during the school year better predict changes in decoding abilities?
3. Does the rate of feature use or mere presence of an MAE form in an individual's repertoire better predict decoding?
4. Are certain types of differences, as reflected in subscores, more useful at predicting changes in decoding abilities? More specifically, given the close relationship between phonology and decoding, are phonological differences more predictive of changes in decoding than grammatical differences? Additionally, if forms like third singular verbal -*s* are more indicative of a shift to MAE, will their usage be especially predictive of differences in decoding?
5. Does the addition of a secondary sentence repetition task explain differences in children's developing decoding abilities over and above what can be observed from DELV-based measures?

Methods

Participants

The participants were 296 kindergarten and 260 first grade children from 12 elementary schools in the Baltimore City Public Schools. All schools had a minimum of 89% African American students (mean=96%) and more than 89% of students eligible for the National School Lunch Program (mean=94%). All students were participating in a larger study designed to evaluate the efficacy of a dialect-shifting curriculum (Edwards, 2019). Only students who did not have an

Individualized Education Program (IEP) were included in study; 14 students with IEPs were tested but removed from analysis. A total of 69 students were excluded due to absence or transfer at the second of the two testing points. Since model comparison values are only valid for models that have been fit to the same data, an additional 8 participants were excluded due to a lack of scorable items for a DELV subscore, withdrawal of assent on an assessment, or failure to establish a ceiling score on a Basic Reading Cluster assessment due to experimenter error. Thus, the present analysis used data from a total of 475 students (241 kindergarteners, 234 first graders). At baseline, the mean age was 5:8 (S.D.=5 months) for kindergarteners and 6:8 for first graders (S.D.=4 months), and at post, the mean age was 6:2 (S.D.=5 months) for kindergarteners and 7:2 (S.D.=4 months) for first graders.

Procedure

All students were taken out of class for one hour of testing near the beginning of the school year (October) and a second hour of testing near the end of the school year (April-May). There were approximately six months between the first and second testing period. Students were tested individually and received the following assessments: (1) Part 1 of the *Diagnostic Evaluation of Language Variation-Screening Test* (DELV-ST, Seymour et al., 2003); (2) the *Dialect Assessment Battery* (DAB, Craig, 2014); and (3) the Basic Reading Cluster (Word Attack and Letter-Word Identification subtests) from the *Woodcock Johnson Achievement Test, 4th edition* (WJIV; Schrank, Mather, & McGrew, 2014).

The DELV-ST Part 1 includes 15 items that are contrastive between AAE and MAE (see examples in Table 2). Five of these items focus on phonological differences between the two dialects (“DELV-Phon”), and the remaining ten items focus on differences in subject-verb agreement between the dialects. Six of these items (“DELV-Irreg”) test the irregular subject-verb agreement patterns of the verbs *have/has*, *don't/doesn't*, and *was/were*, and four of these items ("DELV-3sg") test use of regular verbal *-s* with third person subjects (e.g., *The girl sleeps*). The DELV-ST provides a criterion score of "strong variation from MAE," "some variation from MAE," or "no variation from MAE." At baseline, 83% of participants were in the "strong variation" category, 4% were in the "some variation" category, and 13% were in the "no variation" category. At post, 80% were in the "strong variation" category, 4% were in the "some variation" category, and 15% were in the "no variation" category.

The DAB is a non-standardized test that is designed to be used with *ToggleTalk*®, a dialect shifting curriculum supplement for kindergarten and first grade students (Craig, 2014). We administered a form of the DAB that was adapted to target one feature per item, and sentences were recorded by the same set of four individuals who speak both AAE and MAE. The DAB is composed of three subtests. Part 1, Elicited Imitation, assesses children's ability to repeat sentences produced in MAE. Part 2 assesses Dialect Recognition; students are asked to state whether each sentence is produced in AAE (informal/home talk) or MAE (formal/school talk). Part 3, Translation/Reformulation, asks children to translate sentences from AAE to MAE. All three sections include 12 items: two sentences with plural forms, two sentences with past tense, three sentences with a copula, three sentences which focus on subject-verb agreement (two sentences with third-singular /s/ and one sentence with plural subject and "were"), and two sentences with possessive /s/. Only Elicited Imitation (Part 1, "DAB-EI") and Translation (Part 3, "DAB-TR") were used in the present analysis.

The Basic Reading cluster of the WJIV assesses children's ability to read words (Letter Word Identification subtest) and nonwords (Word Attack subtest). This measure provides both standard scores with a standardized mean of 100 and a standard deviation of 15 and *W*-scores, which are linear raw scores. The Basic Reading standard and *W*-scores are the mean of Letter Word Identification and Word Attack scores. We used the Basic Reading *W*-scores in our analysis. Table 1 provides mean scores for all assessment measures used in the modeling.

Insert Table 1 about here

Analysis

Dialect variation (DVAR) score. For the frequency-based approach, we calculated a *dialect variation* score (DVAR, a type of NMFD measure) from the DELV-ST. This score was computed by dividing the total number of items that varied from MAE from the total number of scorable items and multiplying by 100; a child who uses a nonmainstream form on every item will receive a score of 100 (Terry et al., 2010; Terry & Connor, 2012; Terry et al., 2012). Additionally, we calculated three DVAR subscores corresponding to phonological differences (DVAR-Phon), irregular subject-verb agreement (DVAR-Irreg), and regular subject-verb agreement (DVAR-3sg). These DVAR scores were used in this analysis.

The appropriateness of our selection of subscores is also supported by a confirmatory factor analysis. Confirmatory factor analysis with oblimin rotation (i.e., allowing for correlated factors) was performed using Mplus. Due to the discrete scale of the items, a mean and variance adjusted weighted least squares estimation approach was used to extract factors (Liang & Yang, 2014). Each item was coded as a binary variable, where 1 corresponded to use of an AAE feature that is not grammatical in MAE and 0 corresponded to an MAE-compatible utterance; all other responses were treated as missing data. A two-factor model corresponding to phonological (items 1-5) and morphological features (items 6-15) did not provide good data-model fit

($X^2(91)=4027.76$, $p<0.001$, RMSEA=0.20, CFI=0.63, SRMR=0.24). However, satisfactory data-model fit was obtained with a simple three-factor structure, where morphological features were split into regular and irregular subject-verb agreement (i.e., factors corresponding to DELV-Phon, DELV-Irreg, and DELV-3sg; $X^2(87)=449.11$, $p<0.001$, RMSEA=0.063, CFI=0.97, SRMR=0.07). In order to facilitate comparison with DAB scores, we generated factor scores from a five-factor structure, which also provided good data-model fit ($X^2(692)=1375.89$, $p<0.001$, RMSEA=0.03, CFI=0.96, SRMR=0.09), with additional factors corresponding to AAE use on the Elicited Imitation and Translation components of the Dialect Assessment Battery (see Table 2). No model included cross-loadings. Regression analyses using factor scores were qualitatively similar to those using DVAR subscores with DAB total scores, so only the results from DVAR subscores and DAB total scores are reported here. Analyses with factor scores can be found in S1.

Insert Table 2 about here

DAB score. Each item of the DAB received a score of 2, 1, or 0; where “2” corresponded to MAE use, “1” corresponded to partial credit, and “0” corresponded to any other response, including responses that involved a nonmainstream feature, since the assessment explicitly prompts the use of MAE. Items received 2 points if the child produced the exact sentence of MAE that was targeted, with credit awarded if proper names were changed in the child’s utterance. For both Elicited Imitation and Translation, children received 1 point if their response was grammatical in MAE but the sentence was modified. For Translation, children could also receive 1 point if they produced the targeted MAE feature but another portion of the sentence was changed, even if this change made the sentence ungrammatical in MAE. A total score out of 24 was calculated for each subsection.

Repertoire. To measure which phonological and morphological features that differentiate AAE and MAE were in a child's repertoire, we measured repertoire as a binary variable where a score of 1 indicated that the child had used at least one form compatible with MAE, and a score of 0 indicated that a child had not used an MAE form at the time point in question. This was calculated for the three subcomponents of the DELV (DELV-Phon, DELV-Irreg, and DELV-3sg), as well as the following features on DAB Translation: overt present-tense copula (3 items); overt past tense marking (2 items); overt plural marking (2 items); and overt possessive marking (2 items).[1]

Repertoire values can be interpreted as a measurement of whether a child ever uses a given form that is part of MAE, regardless of whether they sometimes (or even primarily) use a different form. However, we should note that our scores are not derived from assessments that target repertoire. The DELV-ST uses sentence completion to maximize the elicitation of nonmainstream forms; it is designed to help clinicians identify if a child might speak a nonmainstream variety of English. The DAB-Translation, on the other hand, explicitly prompts children to use "school language." Success in this task pre-supposes presence of the MAE-compatible form in the child's repertoire, but the task requires further metalinguistic skill and choices of self-expression. Thus, for both tasks, it is possible for an MAE-compatible form to be in a child's repertoire but not be elicited; however, it would not make sense for a form to be observed if it is not in a child's repertoire. We are testing whether this measure has predictive value, despite its limitations.

[1] Due to an oversight in stimulus preparation, children could provide a valid translation of the subject-verb agreement items without using verbal *-s*, so these items were excluded.

Statistical analysis. Unless otherwise noted, we used linear mixed-effects regression models (Fitzmaurice, Laird & Ware, 2011) to test the predictive value of each dialect measure. Models were fit using the lme4 package (version 1.1-21; Bates, Maechler, Bolker & Walker, 2015) in R (version 3.6.1) using restricted maximum likelihood estimation. We used the package lmerTest (version 3.1-0; Kuznetsova, Brokhoff & Christensen, 2017) to calculate *p*-values for model coefficients using Satterthwaite's method. Standardized parameter estimates ($\widehat{\beta^*}$) are provided as a measure of effect size.

Results

In the sections below, we describe the results of our analyses with the measures of AAE usage described above: DVAR scores, DAB scores, and repertoire values. We used these different measures of AAE usage for two purposes: (1) to describe change in dialect use from the beginning to the end of the school year; and (2) to predict decoding skills at the end of the school year. Figure 1 shows changes in dialect usage from the beginning to the end of the school year for all modeled measures of dialect usage. Model results for significant and marginally significant effects are provided in this section; full model results are provided in S1.

Relationships Among Dialect Measures

A total of 13 dialect measures were generated from the items of the DELV-ST and DAB. These corresponded to a composite DVAR score from all 15 DELV items; DVAR-Phonology, DVAR-3sg, and DVAR-Irreg; total DAB scores for Elicited Imitation and Translation; repertoire scores generated from DELV-Phon, DELV-3sg, and DELV-Irreg; and repertoire scores generated from the DAB for overt present-tense copula, overt past tense morphology, overt plural morphology, and overt possessive morphology. Correlations among all of these measures at baseline, post, and between baseline and post can be found in S2, and a correlation matrix of DVAR scores, DAB scores, and repertoire scores at baseline is provided in Table 3.

Insert Table 3 about here

For each measure, scores at baseline were significantly correlated with scores at post at the α=0.05 level, and all of these correlations were positive. Additionally, each variable was significantly correlated with DVAR-Composite at both baseline and post, with the following exceptions: DAB-Rep-Plural at baseline with DVAR-Composite at baseline (r=-0.08, p=0.07) and post (r=-0.04, p=0.34), and DAB-Rep-Past at baseline with DVAR-Composite at post (r=-0.08, p=0.08). Because of this, DAB-Rep-Plural was not included in subsequent models.

Changes in Dialect Measures Over Time

For each of our measures, we confirmed the widely-observed trend of decreases in NMFD throughout early school years (e.g., Terry et al., 2010). We used linear mixed-effects models to measure change in each score. Each score was modeled separately, with fixed effects of time point (fall or spring), grade level, and their interaction, as well as participant- and classroom-level random intercepts and classroom-by-time point random slopes.

DVAR and DAB. For DVAR-Composite, there was a significant effect of time point ($\widehat{\beta^*}$=-0.18, S.E.=1.38, t(42.06)=-3.01, p=0.004), indicating a decrease in NMFD for kindergarteners over the course of the school year. There was a significant effect of Grade ($\widehat{\beta^*}$=-0.28, t(39.65)=-2.29, p=0.027), indicating that first graders at baseline have lower DVAR scores than kindergarteners at baseline. Finally, there was a significant interaction ($\widehat{\beta^*}$=-0.17, t(41.8)=-2.06, p=0.045), indicating that the decrease in DVAR between fall and spring was more pronounced for first graders, relative to kindergarteners. For the DVAR-Phon subscore, there was a significant effect of grade ($\widehat{\beta^*}$=-0.22, *t*(43.49)=-2.27, *p*=0.028) and a significant time point by grade level interaction ($\widehat{\beta^*}$=-0.25, *t*(39.78)=-2.22, *p*=0.032), suggesting that DVAR phonology scores decrease, but only during first grade. For the DVAR irregular subscore, there was a significant effect of time point ($\widehat{\beta^*}$=-0.21, *t*(41.04)=-3.83, *p*<0.001) and a marginal effect of grade ($\widehat{\beta^*}$=-0.25, *t*(37.52)=-1.93, *p*=0.061), but no time point by grade interaction; this indicates a significant increase in the use of *has*, *doesn't*, and *were* over the course of the school year, with a potentially higher starting point in grade 1. There were no significant terms for the DVAR-3sg subscore; this indicates that there was no increase in use of the third person singular from kindergarten to first grade or from the beginning to the end of the school year.

For overall DAB-EI, there was a significant effect of time point ($\widehat{\beta^*}$=0.33, *t*(44.55)=5.54, *p*<0.001) and grade ($\widehat{\beta^*}$=0.46, *t*(42.86)=3.94, *p*<0.001), but not an interaction, indicating that usage of MAE in sentence repetition increases over the course of the school year and between kindergarten and first grade. For overall DAB-TR, there was also a significant effect of time point ($\widehat{\beta^*}$=0.29, *t*(125.54)=4.2, *p*<0.001) and grade ($\widehat{\beta^*}$=0.45, *t*(64.74)=4.72, *p*<0.001), as well as an interaction ($\widehat{\beta^*}$=0.32, *t*(124.68)=3.31, *p*=0.001), indicating that children's ability to translate sentences from AAE into MAE increases over the course of the school year and between kindergarten and first grade, and this effect is more pronounced in first grade.

Insert Figure 1 about here

Repertoire. We ran mixed-effects logistic regression models, which are appropriate for predicting binary-coded data, with fixed effects of grade and time point and their interaction, as well as participant- and classroom-level random intercepts, using the glmer function of lme4. A separate model was fit for each repertoire score. For overt copula usage, there was a significant fixed effect of time point ($\widehat{\beta^*}$=0.91, z=4.50, p<0.001) and grade level ($\widehat{\beta^*}$=0.90, z=3.93, p<0.001), indicating that children were more likely to have overt copula in their repertoire in the spring relative to the fall and in first grade relative to kindergarten; for overt possessive usage, there was a significant interaction term ($\widehat{\beta^*}$=1.02, z=2.97, p=0.003), meaning that overt possessive was more likely to be in a child's repertoire at spring testing in grade 1, relative to any other time point. No other terms were significant.

Predicting Decoding from Dialect Measures

We ran two sets of models to examine the relationship between NMAE use and reading scores. In one set of models, we examined whether change in NMAE use across the school year was a significant predictor of reading scores at the end of the school year. These models tested the claim that being successful at the linguistic and metalinguistic demands inherent in learning to dialect shift is associated with learning to read (e.g., Terry & Scarborough, 2011). In the second set of models, we examined whether baseline NMAE scores were significant predictors of reading at the end of the school year. These models tested the claim that learning to decode is more difficult for children with higher rates of NMAE use, probably because of the greater mismatch between their native dialect and the written form (e.g., Labov, 1995). Models did not converge or had a singular fit using classroom-level random slopes for the relationship between baseline (fall) scores and post (spring) scores, so we simplified our random effects structure to include classroom-level random intercepts only. The intraclass correlation was 0.41 for the unconditional model. Additionally, inclusion of grade level in models predicting reading did not significantly improve model fit, so the term was dropped.

Dialect variation scores (DELV-Screening Test).

DVAR growth predicting decoding. We used linear mixed-effects regression models to predict Basic Reading *W*-scores in the spring, with fixed effects of fall *W*-scores and the difference between fall and spring DVAR scores. For the model with overall DVAR change, there was a significant effect of fall Basic Reading *W*-scores ($\widehat{\beta^*}$=0.85, t(376.32)=36.27, p<0.001), indicating that students with higher Basic Reading scores in the fall had higher Basic Reading scores in the spring, and there was a significant effect of change in DVAR ($\widehat{\beta^*}$=-0.08, t(466.96)=-3.97, p<0.001), indicating that, controlling for Basic Reading score in the fall, children had higher Basic Reading scores in the spring as their NMFD decreased. Addition of DAB-EI and DAB-TR score changes marginally improved model fit (X^2(2)-5.43, p=0.066), driven by a marginal effect of DAB-TR ($\widehat{\beta^*}$=0.04, t(445.55)=1.86, p=0.064).

We fit a separate model using the three DVAR subscores as independent predictors. There was a significant effect of baseline reading score ($\widehat{\beta^*}$=0.85, t(373.4)=36.19, p<0.001), DVAR phonology subscore change ($\widehat{\beta^*}$=-0.06, t(461.58)=-2.93, p=0.004), and DVAR irregular subscore change ($\widehat{\beta^*}$=-0.05, t(456.4)=-2.54, p=0.011), but not DVAR-3sg subscore change. Again, addition of DAB-EI and DAB-TR score changes marginally improved model fit (X^2(2)-5.43, p=0.066), driven by a marginal effect of DAB-TR ($\widehat{\beta^*}$=0.04, t(444.32)=1.77, p=0.078).

DVAR baseline predicting decoding. Next, we used baseline DVAR scores instead of change in DVAR to predict spring reading scores, controlling for baseline reading score, with a classroom-level random intercept. For the model using DVAR-Composite baseline, there was a significant effect of baseline Reading score ($\widehat{\beta^*}$=0.84, t(406.47)=33.48, p<0.001) and baseline DVAR-Composite ($\widehat{\beta^*}$=-0.05, t(466.25)=-2.11, p=0.036). Addition of DAB-EI and DAB-TR baseline significantly improved model fit (X^2(2)=22.33, p<0.001); in the full model, DVAR-

Composite was no longer significant, but DAB-EI baseline was ($\widehat{\beta^*}$=0.10, $t(46.16)=4.23$, $p<0.001$). For the model using DVAR baseline subscores, no subscore was significant, though the addition of DAB-EI and DAB-TR again significantly improved model fit ($X^2(2)=21.90$, $p<0.001$), driven by a significant DAB-EI term ($\widehat{\beta^*}$=0.10, $t(457.96)=4.17$, $p<0.001$).

Repertoire. Given the exploratory nature of our repertoire scores, we used an incremental model building procedure, starting with a null model with a fixed effect of baseline reading score and a classroom-level random intercept. We then added the three repertoire values derived from the DELV (DELV-Phon, DELV-3sg, DELV-Irreg) for model comparison, then additionally included the two DAB-based repertoire measures (overt copula, overt possessive).

Changes in repertoire predicting decoding. To measure whether change in repertoire predicts changes in reading, we modeled Basic Reading *W*-score in the spring with fixed effects of baseline Basic Reading score and the change in each feature in the child's repertoire, where 1 indicated that the feature was added over the course of the year and 0 means it was not. In the null model, Basic Reading score was significant ($\widehat{\beta^*}$=0.85, $t(380.5)=36.04$, $p<0.001$), but the addition of DELV repertoire scores ($X^2(3)=1.88$, $p=0.598$) and DAB repertoire scores ($X^2(2)=0.90$, $p=0.637$) did not improve model fit.

Baseline repertoire predicting decoding. Addition of baseline DELV repertoire scores to the null model marginally improved fit ($X^2(3)=6.49$, $p=0.090$), and addition of DAB repertoire scores significantly improved model fit ($X^2(2)=14.37$, $p<0.001$). In the full model, there was a significant fixed effect of baseline reading score ($\widehat{\beta^*}$=0.83, $t(392.83)=33.97$, $p<0.001$), as well as overt copula usage ($\widehat{\beta^*}$=0.08, $t(448.06)=3.56$, $p<0.001$), and a marginal effect of DELV-Irreg ($\widehat{\beta^*}$=0.04, $t(458.93)=1.78$, $p=0.075$), but no other term. This indicates that controlling for baseline Basic Reading score, children whose repertoire included overt copula in the present tense and (possibly) MAE-compatible agreement on irregular verbs had significantly higher Basic Reading scores in the spring.

Model Comparison

Akaike information criterion (AIC; Akaike, 1974) values for each model of Basic Reading score are provided in Table 4. AIC values provide a measure of model fit that rewards parsimonious, good data-model fit and penalizes overparameterized models (Anderson, 2008). In other words, models are rewarded when predictors explain variance in the outcome measure, but they are penalized for the number of predictors they use. In contrast to statistical tests that compare two nested models, the AIC is a relative fit measure used descriptively, in which the model with the lowest AIC value is considered the best-fitting model, and models with a difference in AIC value of more than four relative to this best-fitting model are considered to have much weaker support. We used AIC_C values, which correct for smaller sample sizes (Anderson, 2008). We refitted the models using maximum likelihood estimation prior to the calculation of AIC_C values. In the present analysis, the DVAR-Composite (plus DAB) approach results in the best-fitting model when predicting decoding in the spring from dialect scores in the fall, with DVAR subscores (plus DAB) also having some empirical support. This approach is also best overall. Of the models that use changes in dialect scores to predict decoding, the DVAR measures are best.

Insert Table 4 about here

Discussion

Regardless of measurement type, we confirmed the widely-reported trend of decreased AAE use and increased MAE use over the course of the school year and between kindergarten and first grade. This was true for both grade levels, but it was more pronounced in first grade for some measures. Given this initial validation of our measures, we return now to our research questions.

First, we found that a three-factor structure provides satisfactory model fit for Part 1 of the DELV-ST, indicating three clusters of items: phonological items, items with regular subject-verb agreement, and items with irregular subject-verb agreement. Irregular subject-verb agreement spanned multiple verbs (*don't* and *haven't* with third person singular subjects, *was* with plural subjects).

Second, models predicting decoding scores from baseline dialect measures provided better fit than models predicting decoding scores from change in dialect measures. However, this overall finding had a complex relationship with individual measures. Grammatical differences, were only significant in baseline dialect models, with the exception of DVAR-Irreg subscores. On the other hand, DVAR Phonology subscore (based on five items) was significant for models that used change in dialect as predictors, but not for models that used baseline dialect measures. This might indicate that knowledge of MAE grammar is a resource that children can draw upon as they learn to read, and time with this resource is necessary for differences to be observed. Phonology, on the other hand, is directly tied to decoding such that changes in one are predictive of changes in the other. Further exploration of this potential distinction could inform future research on literacy interventions. A curriculum that focuses on phonology might have an immediate impact on decoding, while a grammatical one might require additional time before effects are observed.

Models using repertoire scores had poorer data-model fit than models using NMFD, but one of these models did yield a significant result for overt copula usage. We did not observe any hypothesized differences between grammatical feature types. If verbal *-s* is not part of the AAE grammar, we might predict that usage of verbal -*s* would be a particularly powerful indicator of knowledge of MAE and would therefore predict reading outcomes. However, we did not observe this. One possible explanation for this is that there was no increase in usage of verbal -*s* between kindergarten and first grade or between the baseline and the end of the school year. This result is consistent with other research showing that non-overt marking of third person singular shows minimal change from kindergarten to fifth grade (Craig & Washington, 2004; Newkirk-Turner & Green, 2016). Instead, usage of overt copulas was significant in the model predicting reading from baseline repertoire, even though overt copulas are available in both AAE and MAE.

Finally, addition of the DAB did provide predictive value beyond the DELV-ST. The Elicited Imitation subtest of the DAB at baseline predicted decoding in the spring, and this proved to be a stronger predictor than any DELV measure when both were included in the same model. This task is different from the DELV in that it uses MAE forms in the prompts, representing a wider variety of features, and the task is repetition rather than filling in a blank. It is unclear which of these differences was most important, but it is clear that even a brief, highly structured task in addition to the DELV can be useful when the goal is to characterize the language of a child with typical development during early literacy instruction. The elicited imitation task of Charity et al. (2004), which used a picture book context, may have even stronger predictive value than the simple sentence imitation task on the DAB. A comparison of means and standard deviations from the Charity et al. task and our task shows more variability in performance and less of a ceiling effect for Charity et al., suggesting that the picture book context results in a more sensitive measure. We speculate that the picture book context promotes deeper linguistic processing instead of reliance on verbal working memory.

One distinction that did emerge in this work is the difference between agreement in irregular verbs and in other verbs. The relevant DELV-ST items loaded onto separate but correlated factors, and the DVAR-Irreg measure was the only grammatical measure that was significant in a growth model. This could be partially driven by the number of items (6 for DELV-Irreg vs. 4 for DELV-3sg) leading to a reduction in measurement error for the DVAR-Irreg measure. However, it is also plausible that children learn the irregular agreement patterns without learning to use verbal *-s* on regular verbs. This aligns with early findings by Oetting and McDonald (2002), who found that there were more NMAE-speaking children who used zero-marking on regular third person singular verbs than who used nonmainstream subject-verb agreement with *be* and *don't.* Given the frequency of these forms, the relationship between knowledge of MAE agreement patterns for irregular verbs and reading could conceivably operate in either causal direction. Stronger readers might learn these forms from their experiences with texts, and knowledge of these frequent forms could lead to greater facility with decoding.

While we provided multiple ways of analyzing DELV-ST and DAB data, we are limited to these two assessments in our dialect measures. Both provide highly structured elicitation contexts, and previous work has made it clear that children's dialect usage differs depending on the elicitation context (e.g., Craig et al., 2014; Renn and Terry, 2009). More open-ended narrative tasks could provide useful comparison data and potentially elicit more nonmainstream components of a child's repertoire, but such tasks are more difficult to administer and score on a large scale. Such tasks also provide an opportunity to measure style shifting across contexts within a given time point, rather than confounding changes in dialect and development.

Additionally, our reading measures are limited to measures of decoding. Specifically, the Basic Reading composite score that we used is calculated from two subtests that evaluate the reading of words and non-words in isolation. While these are appropriate reading measures for children at this stage of schooling, it is possible that dialect differences play a distinct role in passage reading (e.g., Terry et al., 2016). That is, grammatical differences between dialects may be more important for passage comprehension than for decoding, since grammatical differences such as agreement morphemes are more likely to appear in a passage than in isolated words.

We are also limited by our relatively homogeneous sample. By design, our participants attended schools where students were predominantly African American and from low-SES families, and all of these schools were part of the same district. Moreover, the majority of participants (89%) showed at least some variation from MAE as measured by the DELV-ST. Terry et al. (2012) found significant effects of race, school SES, and race-by-SES interaction in predicting change in DVAR scores, so more research will be necessary to determine the degree to which our results generalize to speakers of other nonmainstream dialects of English and to other school settings. It is plausible that different types of experiences with variation would be better captured by different measures, which is important to note when comparing studies that use different populations of speakers.

Further research will be necessary to confirm these exploratory findings. This will involve continued honing of our measurements of dialect differences. Future studies could elicit larger numbers of tokens per feature and systematically vary the elicitation context to include sentence repetition, sentence completion, and open-ended narrative. This would allow us to more clearly determine which combination(s) of tasks and dialect differences are most predictive of changes in measures of reading. This process should be repeated across multiple age ranges to reflect children's evolving dialect usage. Factor analysis played only a limited role in the present study, but it is a promising tool for future research. Ideally, future work would use structural equation modeling not only for measures of dialect but also for studying the relationship between those measures and reading scores in order to fully account for measurement error in the measured variables (e.g., Johnson, Terry, Connor, and Thomas-Tate, 2017; Bühler et al., 2018).

As this research progresses, clinicians are faced with the challenging task of supporting speakers of AAE despite having a relatively limited set of tools. One important step is to characterize each child's linguistic repertoire. We have seen that the DELV-ST provides a useful starting point, and it can be made more informative by grouping the items into phonology, regular subject-verb agreement, and irregular subject-verb agreement. Though it might be ideal to use open-form narrative tasks in a variety of settings, even a simple sentence repetition task like the DAB Elicited Imitation can be helpful. As noted above, a sentence imitation task that incorporates a storybook or picture description as part of the paradigm (e.g., Charity et al., 2004) may result in greater semantic encoding of the sentences and thus elicit a representative range of nonmainstream forms. More broadly, it is important to think of any measure of nonmainstream form density as only a starting point for understanding a child’s language and anticipating any educational challenges from linguistic differences. The next step is providing targeted support. Our results provide tentative support for the idea of focusing on areas of variable overlap between MAE and AAE, such as overt copulas and irregular subject-verb agreement. This allows children to draw upon their existing linguistic knowledge as they learn to read in a less familiar dialect.

Acknowledgements

This research was supported in part by IES grant #R305A17013 to Jan Edwards and NSF grant #1449815 to Colin Phillips. We are very grateful to all of the children who participated in this study, their families who provided consent, and all of the amazing teachers and principals in the Baltimore City Public School System whom we have worked with. We also thank Ebony Terrell Shockley, Rebecca Silverman, Tatiana Thonesavanh and other members of the Learning to Talk Lab at the University of Maryland for their many contributions to this research program.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716-723. https://doi.org/10.1109/TAC.1974.1100705

Anderson, D. R. (2008). *Model based inference in the life sciences: a primer on evidence*. (Vol. 1). New York: Springer.

Auer, P. (2005). Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. In *Perspectives on Variation: Sociolinguistic, Historical, Comparative*, (pp. 7–42).

Barrière, I., Kresh, S., Aharodnik, K., Legendre, G., & Nazzi, T. (2019). The comprehension of 3rd person singular –s by NYC English-speaking preschoolers. In In T. Ionin & Rispoli, M. (Eds.). *Selected Proceedings of the 7th Generative Approaches to Language Acquisition - North America Conference* (pp. 7-33). Philadelphia, PA: John Benjamins.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01.

Baugh, J. (1990). A survey of the suffix /-s/ analyses in Black English. In Edmondson, J. A., Feagin, C., & Mühlhäusler, P., (Eds.), *Development and Diversity: Language Variation across Time and Space: A Festschrift for Charles-James N. Bailey* (pp. 297–307). Dallas, Texas: The Summer Institute of Linguistics.

Beyer, T. & Hudson Kam, C. L. (2012). First and second graders' interpretation of Standard American English morphology across varieties of English. *First Language*, *32*(3), 365-384. https://doi.org/10.1177/0142723711427618

Brown, R. (1973). *A first language: The early stages*. Harvard U. Press.

Bühler, J. C., von Oertzen, T., McBride, C. A., Stoll, S., & Maurer, U. (2018). Influence of dialect use on early reading and spelling acquisition in German-speaking children in Grade 1. *Journal of Cognitive Psychology*, *30*(3), 336-360. https://doi.org/10.1080/20445911.2018.1444614

Champion, T. B., Rosa-Lugo, L. I., Rivers, K. O., & McCabe, A. (2010). A Preliminary Investigation of Second-and Fourth-Grade African American Students' Performance on the Gray Oral Reading Test—Fourth Edition. *Topics in Language Disorders*, *30*(2), 145-153. Https://doi.org/10.1097/TLD.0b013e3181e04056

Charity, A. H., Scarborough, H. S., & Griffin, D. M. (2004). Familiarity with School English in African American children and its relation to early reading achievement. *Child Development*, *75*(5), 1340-1356. https://doi.org/10.1111/j.1467-8624.2004.00744.x

Cleveland, L. H. & Oetting, J. B. (2013). Children’s marking of verbal s by nonmainstream English dialect and clinical status. *American Journal of Speech-Language Pathology*, *22*(4), 604-614. https://doi.org/10.1044/1058-0360(2013/12-0122)

Connor, C. M. & Craig, H. K. (2006). African American preschoolers’ language, emergent literacy skills, and use of African American English: a complex relation. *Journal of Speech, Language, and Hearing Research*, *49*(4), 771-792. https://doi.org/10.1044/1092-4388(2006/055)

Craig, H. K., & Washington, J. A. (2000). An assessment battery for identifying language impairments in African American children. *Journal of Speech, Language, and Hearing Research, 43*(2), 366-379. https://doi.org/10.1044/jslhr.4302.366

Craig, H. K., Kolenic, G. E. & Hensel, S. L. (2014). African American English speaking students: A longitudinal examination of style shifting from kindergarten through second grade. *Journal of Speech, Language, and Hearing Research, 57*(1)*,* 143-157. https://doi.org/10.1044/1092-4388(2013/12-0157)

Craig, H. K. (2014). *Toggle Talk*®. Sun Prairie, WI: Ventris Learning.

Craig, H. K. (2016). *African American English and the Achievement Gap*. New York: Routledge.

Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics*, *12*(4), 453-476. https://doi.org/10.1111/j.1467-9841.2008.00374.x

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, *41*(1), 87-100. https://doi.org/10.1146/annurev-anthro-092611-145828

Edwards, J. (2019). Dialect mismatch and learning to read: Research to practice. Plenary talk presented at the 44th Annual Boston University Conference on Language Development, Boston, MA, 7-10 November.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied longitudinal analysis*. Hoboken, NJ: John Wiley & Sons.

Giles, H. & Ogay, T. (2007). Communication Accommodation Theory. In Whaley, B. B. and Samter, W., (Eds.), *Explaining communication: Contemporary theories and exemplars*, (pp. 293-310). New York: Routledge. https://doi.org/10.1002/9781118766804.wbiect056

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *Remedial and special education*, *7*(1), 6-1. https://doi.org/10.1177/074193258600700104

Green, L. J. (2011). *Language and the African American Child*. Cambridge University Press.

Horton–Ikard, Ramonda, and Susan Ellis Weismer. "Distinguishing African American English from developmental errors in the language production of toddlers." *Applied Psycholinguistics* 26.4 (2005): 597-62. https://doi.org/10.1017/S0142716405050320

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and writing*, *2*(2), 127-16. https://doi.org/10.1007/BF00401799

Johnson, L., Terry, N. P., Connor, C. M., & Thomas-Tate, S. (2017). Effects of dialect awareness instruction on nonmainstream American English speakers. *Reading and Writing*, *30*(9), 2009-2038. https://doi.org/10.1007/s11145-017-9764-y

King, S. (2018). *Exploring social and linguistic diversity across African Americans from Rochester, New York* (Doctoral dissertation, Stanford University).

King, S. (2020). From African American Vernacular English to African American Language: Rethinking the Study of Race and Language in African Americans' Speech. *Annual Review of Linguistics*, *6*, 285-30. https://doi.org/10.1146/annurev-linguistics-011619-030556

Kuznetsova A, Brockhoff PB, Christensen RHB (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." Journal of Statistical Software, 82(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Labov, W. (1995). Can reading failure be reversed: A linguistic approach to the question. In V. Gadsden & D. Wagner, (Eds.), *Literacy among African-American youth: Issues in Learning, Teaching, and Schooling* (pp. 39-68). Cresskill, NJ: Hampton.

Liang, X., & Yang, Y. (2014). An evaluation of WLSMV and Bayesian methods for confirmatory factor analysis with categorical indicators. *International Journal of Quantitative Research in Education*, *2*(1), 17-38. https://doi.org/10.1504/IJQRE.2014.060972

McDonald, J. L., & Oetting, J. B. (2019). Nonword repetition across two dialects of English: Effects of specific language impairment and nonmainstream form density. *Journal of*

Speech, Language, and Hearing Research, *62*(5), 1381-1391. https://doi.org/10.1044/2018_JSLHR-L-18-0253

Newkirk-Turner, B. L., & Green, L. (2016). Third person singular-s and event marking in child African American English. *Linguistic Variation*, *16*(1), 103-130. https://doi.org/10.1075/lv.16.1.05new

Newkirk-Turner, B. L., Oetting, J. B., & Stockman, I. J. (2014). BE, DO, and modal auxiliaries of 3-year-old African American English speakers. *Journal of Speech, Language, and Hearing Research*, *57*(4), 1383-1393. https://doi.org/10.1044/2014_JSLHR-L-13-0063

Oetting, J. B. & McDonald, J. L. (2002). Methods for characterizing participants' nonmainstream dialect use in child language research. *Journal of Speech, Language, and Hearing Research*, *45*(3), 505-518. https://doi.org/10.1044/1092-4388(2002/040)

Oetting, J. B., & Pruitt, S. (2005). Southern African-American English use across groups. *Journal of Multilingual Communication Disorders*, *3*(2), 136-144. https://doi.org/10.1080/14769670400027324

Ogbu, J. U. (1999). Beyond language: Ebonics, proper English, and identity in a Black-American speech community. *American Educational Research Journal*, *36*(2), 147-184. https://doi.org/10.3102/00028312036002147

Renn, J. & Terry, J. M. (2009). Operationalizing style: quantifying the use of style shift in the speech of African American adolescents. *American Speech*, *84*(4), 367-390. https://doi.org/10.1215/00031283-2009-030

Roy, J., Oetting, J. B., & Moland, C. W. (2013). Linguistic constraints on children's overt marking of BE by dialect and age. *Journal of Speech, Language, and Hearing Research*. https://doi.org/10.1044/1092-4388(2012/12-0099)

Schrank, F.A., Mather, N., & McGrew, (2014). *Woodcock Johnson IV Tests of Achievement*, NY: Houghton Mifflin Harcourt.

Seymour, H. N., Roeper, T. W., de Villiers, J., & de Villiers, P. A. (2003). *Diagnostic Evaluation of Language Variation – Screening Test*. San Antonio, TX: Pearson.

Snell, J. (2013). Dialect, interaction and class positioning at school: From deficit to difference to repertoire. *Language and Education*, *27*(2), 110-128. https://doi.org/10.1080/09500782.2012.760584

Terry, J. M., Hendrick, R., Evangelou, E., & Smith, R. L. (2010). Variable dialect switching among African American children: Inferences about working memory. *Lingua*, *120*(10), 2463-2475. https://doi.org/10.1016/j.lingua.2010.04.013

Terry, N. P. & Scarborough, H. S. (2011). The Phonological hypothesis as a valuable frame-work for studying the relation of dialect variation to early reading skills. *Explaining Individual Differences in Reading: Theory and Evidence* (pp. 97-117).

Terry, N. P. & Connor, C. M. (2012). Changing nonmainstream American English use and early reading achievement from kindergarten to first grade. *American Journal of Speech-Language Pathology*, *21*, 78-86. https://doi.org/10.1044/1058-0360(2011/10-0093)

Terry, N. P., Connor, C. M., Thomas-Tate, S., & Love, M. (2010). Examining relationships among dialect variation, literacy skills, and school context in first grade. *Journal of Speech, Language, and Hearing Research*, *53*, 126-146. https://doi.org/10.1044/1092-4388(2009/08-0058)

Terry, N. P., Connor, C. M., Petscher, Y., & Ross Conlin, C. (2012). Dialect variation and reading: is change in nonmainstream American English use related to reading achievement in first and second grades? *Journal of Speech, Language, and Hearing Research*, *55*(1), 55-69. https://doi.org/10.1044/1092-4388(2011/09-0257)

Terry, N. P., Connor, C. M. D., Johnson, L., Stuckey, A., & Tani, N. (2016). Dialect variation, dialect-shifting, and reading comprehension in second grade. *Reading and Writing*, *29*(2), 267-295. https://doi.org/10.1007/s11145-015-9593-9

Van Hofwegen, J., & Wolfram, W. (2017). On the utility of composite indices in longitudinal language study: the case of African American language. In *Panel Studies of Variation and Change* (pp. 89-114). Routledge.

Washington, J. A. and Craig, H. K. (2002). Morphosyntactic forms of African American English used by young children and their caregivers. *Applied Psycholinguistics*, *23*(2), 209-231. https://doi.org/10.1017/S0142716402002035

Wolfram, W. (2007). Sociolinguistic Folklore in the Study of African American English. *Language and Linguistics Compass*, *1*(4), 292-313. https://doi.org/10.1111/j.1749-818X.2007.00016.x

Wyatt, T. The acquisition of African American English copula. In: Kamhi, A.; Pollock, K.; Harris, J., editors. *Communication development and disorders in African American children*. Paul H. Brookes Publishing Co; Baltimore, MD: 1996. p. 95-116.

Supplemental Materials

S1: Summary table of fixed effects and lme4 model specification for each model reported in the text. Models with factor score predictors are also included.

S2: Correlations between each pair of dialect measures at both baseline and post

Table 1. Means (standard deviations in parentheses) for assessment measures.

		Kindergarten		First Grade	
Measure Type	Measure	Baseline (Fall)	Post (Spring)	Baseline (Fall)	Post (Spring)
Woodcock Johnson IV	LetterWordID SS[1]	89.57 (14.42)	92.25 (14.57)	85.95 (15.32)	89.01 (17.14)
	LetterWordID *W*-Score[2]	366.87 (3.23)	394.05 (29.61)	40.80 (32.20)	426.01 (34.99)
	WordAttack SS[1]	91.01 (15.38)	96.18 (15.74)	92.14 (16.56)	96.47 (16.95)
	WordAttack *W*-Score[2]	42.88 (24.29)	444.08 (23.54)	45.31 (24.16)	466.15 (22.35)
	Reading SS[1]	9.39 (14.37)	94.22 (14.59)	88.96 (15.16)	92.74 (16.49)
	Reading *W*-Score[2]	393.87 (26.01)	419.07 (25.40)	425.56 (26.86)	446.08 (27.73)
DVAR	Composite	82.61 (19.90)	78.63 (22.07)	75.98 (22.91)	67.84 (25.87)
	Phon	85.02 (21.42)	83.03 (24.01)	79.23 (26.05)	7.77 (32.02)
	3SG	87.52 (25.45)	86.20 (25.36)	82.73 (28.57)	78.88 (32.79)
	Irreg	76.61 (3.48)	69.55 (34.33)	67.69 (33.97)	56.12 (35.76)
DAB	Elicited Imitation (EI)	16.90 (4.61)	18.31 (4.35)	18.85 (3.86)	19.98 (3.47)
	Translation (TR)	6.49 (4.01)	7.98 (4.65)	8.75 (4.79)	11.87 (5.45)
Repertoire	DELV-Phon	.43 (.50)	.44 (.50)	.52 (.50)	.63 (.48)
	DAB-Copula	.32 (.47)	.52 (.50)	.51 (.50)	.68 (.47)
	DELV-3sg	.25 (.44)	.29 (.46)	.34 (.48)	.38 (.49)
	DELV-Irreg	.51 (.50)	.57 (.50)	.62 (.49)	.74 (.44)
	DAB-Past	.50 (.50)	.51 (.50)	.54 (.50)	.53 (.50)
	DAB-Plural	.30 (.46)	.40 (.49)	.38 (.49)	.56 (.50)
	DAB-Possessive	.19 (.39)	.17 (.37)	.18 (.38)	.32 (.47)

[1]Standardized mean is 100 and one standard deviation is 15. [2]A score of 500 represents normative mean achievement of a ten year old, and one standard deviation is 15.

Table 2. Summary of subscores based on factors used in confirmatory factor analysis

Factor	**Construct**	**Example Item**
DELV-Phon	Usage of AAE phonology	*smooth* pronounced /smuv/
DELV-Irreg	Leveling of subject-verb agreement with *have*, *don't*, and *was*	*The girl* ***have*** *a big kite.* *This girl* ***don't*** *like to swim.* *They* ***was*** *sick*.
DELV-3SG	Zero-marking of regular verbs with 3rd person singular subjects	*The boy always* ***ride*** *a bike.*
DAB-EI	Usage of AAE in a sentence repetition task	Prompt: *She* **is** *on the playground* Response: *She on the playground.*
DAB-TR	Usage of AAE when translating sentences from AAE to MAE	Prompt: *The boys* ***was*** *running.* Translation Goal: *The boys* ***were*** *running.* Response: *The boys* ***was*** *running*.

Table 3. Correlations (*r* values) among DVAR, DAB, and repertoire measures at baseline. For DVAR measures, higher values indicate greater usage of NMAE, and for DAB and Repertoire, higher values indicate greater usage of MAE. More correlation information is available in S2.

		DVAR				DAB		Repertoire				
		Composite	3SG	Irreg.	Phon.	EI	TR	Copula	Past	Poss.	Irreg.	Phon.
DVAR	3SG	.77***										
	Irreg.	.87***	.60***									
	Phon.	.62***	.22***	.27***								
DAB	EI	-.36***	-.28***	-.34***	-.17***							
	TR	-.39***	-.30***	-.35***	-.25***	.31***						
Rep.	Cop.	-.20***	-.13**	-.15***	-.17***	.14**	.54***					
	Past	-.10*	-.06	-.10*	-.05	.18***	.31***	-.01				
	Poss.	-.19***	-.17***	-.19***	-.09	.14**	.38***	.09	.18***			
	Irreg.	-.60***	-.34***	-.75***	-.16***	.28***	.24***	.17***	.06	.10*		
	Phon.	-.47***	-.17***	-.19***	-.79***	.16***	.15**	.15***	.00	.06	.07	
	3SG	-.65***	-.84***	-.50***	-.20***	.23***	.24***	.11*	.05	.15**	.35***	.16***

$^{***}p<0.001$ $^{**}p<0.01$ $^{*}p<0.05$

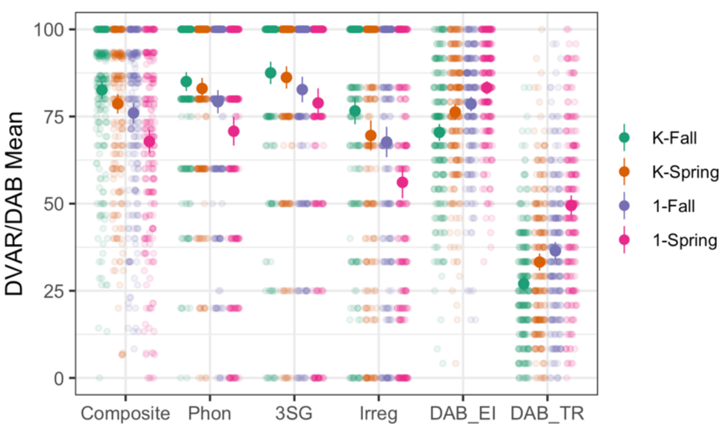
Table 4. Comparison of model fits (lower AIC_C indicates better fit; Δ_i is the difference from the best-fitting model).

	Model	**df**	**AIC_C**	Δ_i
Dialect Change	DVAR (Composite)	7	3826.16	0
	DVAR (Subscores)	9	3828.71	2.55
	Repertoire	9	3848.6	22.43
Dialect Baseline	DVAR (Composite)	7	3820.44	0
	DVAR (Subscores)	9	3824.11	3.68
	Repertoire	9	3830.52	10.08

Figure Captions

Figure 1: Change in dialect usage from the beginning to the end of the school year for kindergarten and first grade students. Error bars represent 95% confidence intervals, and small, semi-transparent points represent individual data points. (a) DVAR scores (higher=greater nonmainstream form density) and DAB total scores (higher=greater MAE use; scores out of 24 have been converted to percentages); (b) Repertoire Scores (of MAE-compatible feature)

(a)



(b)