JSLHR

Research Article

Developing a Longitudinal Scale for Language: Linking Across Developmentally Different Versions of the Same Test

Lee Branum-Martin,[a] Katherine T. Rhodes,[b] Congying Sun,[a] Julie A. Washington,[c] and Mi-Young Webb[c]

Purpose: Many language tests use different versions that are not statistically linked or do not have a developmental scaled score. The current article illustrates the problems of scores that are not linked or equated, followed by a statistical model to derive a developmental scaled score.
Method: Using an accelerated cohort design of 890 students in Grades 1–5, a confirmatory factor model was fit to 6 subtests of the Test of Language Development–Primary and Intermediate: Fourth Edition (Hammill & Newcomer, 2008a, 2008b). The model allowed for linking the subtests to a general factor of language and equating their measurement characteristics across grades and cohorts of children. A sequence of models was fit to evaluate the appropriateness of the linking assumptions.
Results: The models fit well, with reasonable support for the validity of the tests to measure a general factor of language on a longitudinally consistent scale.
Conclusion: Although total and standard scores were problematic for longitudinal relations, the results of the model suggest that language grows in a relatively linear manner among these children, regardless of which set of subtests they received. Researchers and clinicians interested in longitudinal inferences are advised to design research or choose tests that can provide a developmental scaled score.

One of the key challenges in studying the development of language among children is designing tasks that are appropriate to the capabilities of children at different ages. Because the tasks appropriate to younger children differ from those for older children, many language tests offer versions of the same test that are different for different ages or grade levels of the target children. Language tests must be different for basic versus advanced examinees, both for the ethical treatment of participants to reduce frustration and boredom and for valid measurement of the construct (Messick, 1989, 1995). Although such different test versions can represent developmentally and ethically appropriate tasks, we also frequently want to measure growth across developmental phases. Questions of growth over time are central to developmental science (Horn & McArdle, 1992; McArdle & Grimm, 2010).

In order to measure growth across two versions of a test, however, we must define a scale that is common to both. In the same way that we must convert inches into centimeters between an imperial ruler and a metric ruler, we must define a conversion between scores on two different versions of a test (if that conversion is not already provided in the test manual). The purpose of this article is to demonstrate a model for developing a longitudinal metric of growth across two versions of a language test battery.

Unfortunately, such a psychometric link is not given for several frequently used language tests that have different versions for different age levels. Each test gives a standard norm-referenced score, which allows for a comparison to particular age-based mean and standard deviation. Unfortunately, some publishers call such a standard score a *scaled score*. Technically, a *scaled score* is any transformation of a total score (Kolen, 2006; Kolen & Brennan, 1995; Petersen, Kolen, & Hoover, 1989), but in this article, we draw a critical distinction: We will use *standard score* to refer to a norm-based score reflecting standing in a population (e.g., a centile or score referenced to a particular mean and standard deviation of individuals at a particular age—usually cross-sectional),

[a]Department of Psychology, Georgia State University, Atlanta
[b]Department of Developmental Psychology, The Ohio State University, Columbus
[c]Department of Educational Psychology and Special Education, Georgia State University, Atlanta
Correspondence to Lee Branum-Martin: branummartin@gsu.edu

Editor-in-Chief: Sean Redmond
Editor: Lizbeth Finestack
Received September 9, 2018
Revision received December 17, 2018
Accepted January 27, 2019
https://doi.org/10.1044/2019_JSLHR-L-18-0362

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

and we will use *developmental scaled score* or simply *scaled score* to refer to a psychometrically defined longitudinal metric that is consistent over time (e.g., a score developed from item response theory to account for different item properties in a way that is equivalent for different examinees). Unfortunately, some tests call some forms of standard scores *scaled scores*, which goes against the convention we adopt here.

Several popular language tests do not have a developmental scaled score. Developmental scaled scores allow for the measurement of an individual's growth over time using the same, consistent metric, but they require a psychometric link between two versions of a test. If the units are not consistent from one time point to another, then judging gain is difficult, and most crucially for research, statistical models of growth become meaningless and efforts to describe growth can lead to mistakes (Seltzer, Frank, & Bryk, 1994).

An Applied Example: Who Needs a Scaled Score?

To see the conceptual importance of a consistent metric to measure growth (outside statistics and equations), let us consider an example in which we wish to gauge the vocabulary growth rate of students during the elementary school years. We take a sample of 890 students in Grades 1–5 using an accelerated cohort design that spans 2 years (i.e., a student who begins the study in first grade is sampled at Year 1 in first grade and at Year 2 in second grade, whereas a student who begins the study in second grade is sampled at Year 1 in second grade, at Year 2 in third grade, and so on). At each time point, these students are given the Picture Vocabulary subtest of the Test of Language Development–Primary and Intermediate: Fourth Edition (TOLD-P:4 and TOLD-I:4; Hammill & Newcomer, 2008a, 2008b). The test has two versions, namely, Primary and Intermediate, split at 8 years of age, and students are administered the version of the test that is appropriate to their ages. Thus, the average second-grade student may receive the Primary version of the test in Year 1 in second grade and then receive the Intermediate version of the test in Year 2 in third grade (issues of overlap across versions of the test will be discussed later).

Measuring Growth With Total Scores

Figure 1 shows growth lines for students (gray lines), using their total (raw) scores at each time point. The dark line represents the average growth rate across Grades 1–5, ignoring the test version students received (using loess regression to create a moving average for the full sample). The dotted line shows the average growth rate for students given the Primary version of the test, and the dashed line shows the average growth rate for the students given the Intermediate version of the test. The overall, solid line suggests that students are growing on average from Grades 1 to 5. The dashed line for the Intermediate version is quite close to the overall line. The dotted line for the Primary version, however, is quite flat, suggesting that these students are not growing much on average.

This graph would seem to suggest that younger students' vocabularies are not growing—a conclusion that would be problematic, if not fairly disturbing. However, our conclusions about growth are limited, given that a total score on the Primary version of the test may not be equivalent to a total score on the Intermediate version of the test. The growth differences between test versions may have resulted from (a) the tests having different measurement properties (e.g., having unequal difficulty or sensitivity), (b) the younger group of students having a truly flat trajectory, or (c) some mixture of test and group differences that produced the diverging trend lines. However, we cannot be sure that a given total score on one version means the same thing as that score on another version. Without a mathematical link between total scores on the two different test versions, we cannot know whether the longitudinal trends we see are due to child development or due to the test versions being unequal. This is a classic problem of linking or equating different versions or forms of tests (Kolen, 2006; Kolen & Brennan, 1995; Petersen et al., 1989).

Might Standard Scores Save Us?

We might posit that norm-referenced scores could provide a meaningful basis to compare groups of students across test versions. Figure 2 shows a similar growth plot, but for the age-referenced standard scores[1] with a mean of 10 and a standard deviation of 2. A completely average sample, therefore, should have a flat line at a value of 10 units.

In Figure 2, the dotted line for the Primary version suggests that students have a mean of around 10 and are holding relatively steady, perhaps slightly increasing from Grades 1 to 3. The dashed line for the Intermediate version suggests lower performance, but rising, nearly reaching the normative mean at Grade 5. The solid overall line, however, suggests that the sample is falling away from the mean—a widening language gap. Thus, Figure 2 shows an intolerable contradiction of each group holding steady or rising versus the overall sample falling relative to the norm—either of these situations might be true, but not both.

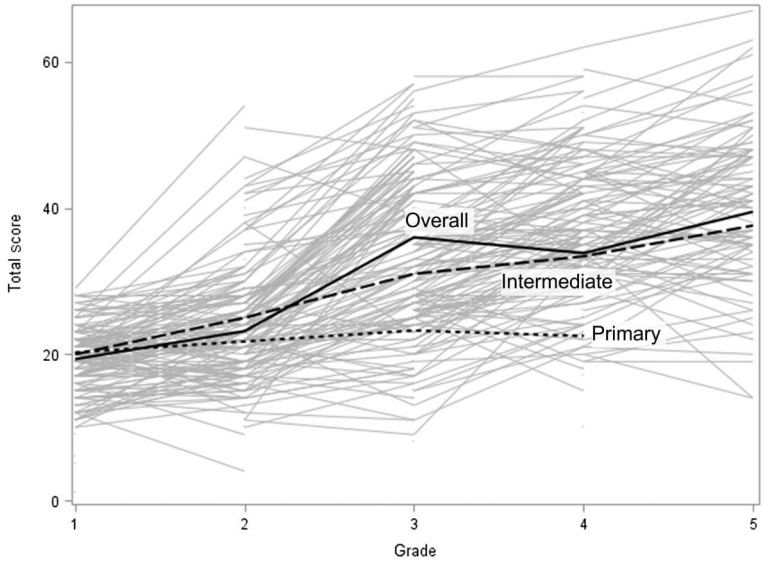
Standard scores, in this case, have not resolved what seem to be differing conclusions about the nature of language growth for the two versions of this test in this sample. As in the case of total scores, we cannot be sure unless we have a clear method to link the scores to a longitudinally consistent metric: We need a developmental scaled score.

This Study

Insofar as language tests have different content at higher versus lower levels of development, longitudinal designs using similar tests will face the same problem. In particular, longitudinal analyses of language development may

[1]Here, we again note that some tests call such a score a "scaled score". Because this score is standardized relative to a particular population mean and standard deviation, we call it a *standard score* and further note that it does not have a longitudinally consistent metric. By our terminology, both the "scaled score" and "standard score" on the TOLD-4 are *standard scores*—they are simply standardized to different units (i.e., *M*/*SD* of 10/2 and 100/15, respectively).

Figure 1. Total score growth trajectories for the Picture Vocabulary subtest. Each student's total score across grades is represented by a gray line. A moving average (loess) regression line is shown for the overall sample (solid line) and for each of the two versions: Primary (dotted line) and Intermediate (dashed line).

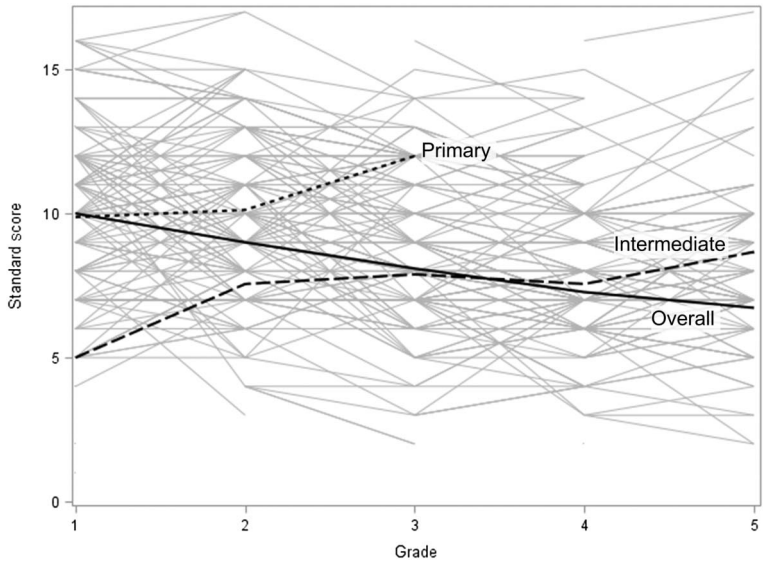


be fraught with artifacts due to the necessary use of different test versions or forms: Some trends or age-based differences may be due to differences in test version (difficulty or sensitivity) instead of due to developmental growth in the students. This study examines a statistical model to equate language test versions across groups and over time.

In order to develop this psychometric link, we can take advantage of the overlapping cohorts in the design and make some measurement assumptions, which we present here in a nontechnical format (readers interested in technical details are referred to the Appendix):

1. Construct validity or unidimensionality: When using multiple subtests (e.g., morphology, vocabulary, and syntax), they measure a single, general ability or factor. This assumption is often implicit in tests that have instructions for a broad or total language score.

Figure 2. Standard score growth trajectories for the Picture Vocabulary subtest of the Test of Language Development–Fourth Edition. Each student's score across time is represented by a gray line. A moving average (loess) regression line is shown for the overall sample (solid line) and for each of the two versions of the Test of Language Development: Primary (dotted line) and Intermediate (dashed line).



2. Cohorts of students are equal: If we measure one group of second graders, they are approximately equivalent on average to other groups of second graders.
3. Measurement properties of the test are equal: A given subtest should measure its intended construct (the factor of language) the same way, regardless of when or to whom it is administered.

Together, we can use these assumptions to build statistical links across cohorts of students, across subtests, and across versions of the test. This statistical model is confirmatory and falsifiable—if it fails to fit adequately, then one or more of our assumptions are wrong. We therefore present a longitudinal confirmatory factor analysis (CFA; Bollen, 1989; Little, 2013; Rock, 1982) in order to jointly scale the two versions of the popular language assessment, the Test of Language Development–Primary and Intermediate: Fourth Edition (TOLD-P:4 and TOLD-I:4, respectively; Hammill & Newcomer, 2008a, 2008b). We examine the extent to which this jointly scaled factor model can be used to measure language development over time on a longitudinally consistent scale. We have previously used scores from a simplified version of this model, which forced this equating without testing it in detail (Washington, Branum-Martin, Sun, & Lee-James, 2018; Washington, Branum-Martin, Lee-James, & Sun, in press). The current study provides the full, empirical evaluation of this scoring approach for the Test of Language Development–Fourth Edition (TOLD-4).

Method

Participants

Participants were drawn from a larger project focused on language, literacy, and dialectal variation, including 890 African American boys and girls in first through fifth grades in a major urban school district in the southeastern United States. Participants were enrolled in seven high-poverty schools where 87%–100% of children qualified for participation in the National School Lunch program, which provides free or reduced priced meals to low-income students. The design was an accelerated cohort design (McArdle & Hamagami, 1991; Meredith & Tisak, 1990; Schaie & Baltes, 1975), in which students in Grades 1–5 were planned to be measured in each of 2 years on three subtests of the TOLD-4. Recruitment and administration procedures have been previously reported (Washington et al., 2018, in press).

Because this is a complex design involving two versions of a test, we will organize the presentation of data around test administration groups (see administration procedure below). Table 1 presents counts of students in each of these administration groups, by grade and test version (Primary or Intermediate) of the TOLD-4 (see Assessment Measures section below). Overall, approximately half of the sample had 2 years of measures, and half had only 1 year observed (see Table 1). Table 2 presents participant characteristics by test administration group. The sample was 50% female, and the children had nonverbal intelligence within normal limits ($M = 97$, $SD = 16$) as measured by the Kaufman Brief Intelligence Test–Second Edition (Kaufman & Kaufman, 2004).

Table 1. Patterns of test administrations in each grade level for the Primary and Intermediate versions of the Test of Language Development–Fourth Edition, with group assignment for the scaling model.

Administration group: grade, version	Grade level 1	2	3	4	5	Pure	Mixed	Total
1: first, Primary	P	P				84	2	86
	—	P				5	0	5
2: first, both versions	P	I				32	3	35
	P	—				126	0	126
	—	I				1	2	3
3: second, both versions		P	I			66	3	69
		P	—			47	0	47
4: second, Intermediate		I	I			40	6	46
		I	—			34	5	39
		—	I			6	0	6
5: third, Intermediate			I	I		95	20	115
			I	—		79	14	93
			—	I		4	1	5
6: fourth, Intermediate				I	I	77	13	90
				I	—	78	7	85
				—	I	15	0	15
Totals						789	76	865

Note. Dashes indicate that the test was not observed in that year (missing). "Pure" indicates students receiving those only test versions. "Mixed" indicates students who tested in multiple versions but had two scores on the version shown (e.g., Intermediate due to age, but dropped back into at least one Primary version subtest). Total $N = 890$ (20 students had no valid scores, and five students had tests outside the given patterns, with insufficient overlap to make a meaningful link across grades for that test version—e.g., Intermediate version in Grade 1 or Primary version in Grade 3). P = Primary version; I = Intermediate version.

Table 2. Demographic characteristics.

Administration group	*n*	Female (%)	Age (years) *M*	Age (years) *SD*	KBIT *M*	KBIT *SD*
1	91	55	6.7	0.4	98	18
2	164	51	6.9	0.5	93	16
3	116	60	7.6	0.2	99	16
4	91	49	8.3	0.4	94	15
5	213	50	8.8	0.5	98	16
6	190	49	9.9	0.5	96	14
Total	865	50	8.3	1.3	97	16

Note. Groups refer to administration groups (see Table 1). KBIT = Kaufman Brief Intelligence Test standard score.

Assessment Measures

Language performance was measured using three selected subtests of the TOLD-P:4 (Hammill & Newcomer, 2008a) and three subtests of the TOLD-I:4 (Hammill & Newcomer, 2008b). The Primary version is designed for use with children between ages 4 years and 8;11 (years; months), and the Intermediate version is designed for children between ages 8 years and 17;11.

In the current study, students aged 8 years and older were administered the TOLD-I:4. In order to avoid floor effects among older students, "drop-back" testing procedures were created so that students who struggled with the Intermediate version of subtests were switched to a Primary version subtest of interest. The administration rules for each of the three subtests are shown in Table 3, with the number of items per test. If students failed to get the specified number of items correct on the Intermediate version, then the corresponding subtest from the TOLD-P:4 was administered instead (Table 2 shows the overlap in test versions due to age).

Semantic Language Development

The Picture Vocabulary subtest on the Primary version consists of 34 items designed to measure semantic listening or children's abilities to understand the meanings of words spoken aloud. Children were presented with an array of four illustrations and selected the illustration that best corresponded to the word spoken by the examiner. Responses were scored as either *correct* (1) or *incorrect* (0). Testing began with Item 1 and ceased when children answered five consecutive items incorrectly. In the current sample, internal consistency of the Primary version's Picture Vocabulary subtest at Time 1 was $\alpha = .77$.

The Picture Vocabulary subtest on the Intermediate version consists of 80 items, again designed to measure semantic listening. Children were presented with a series of nine picture cards, each with an array of six pictures, and asked to select the picture that best reflected a two-word prompt given by the examiner. Testing began with an example item ("monkey see") and proceeded to Picture Card 1 until children had either attempted every item relevant to Picture Card 1 or until they had responded incorrectly (scored "0") for two consecutive items. Testing then proceeded through each of the nine picture cards using the two-item ceiling rule. In the current study, children who were unable to correctly answer the example item were administered the Primary version's Picture Vocabulary subtest (see Table 3). Internal consistency of the Intermediate version's Picture Vocabulary subtest at Time 1 was $\alpha = .89$ for the current sample.

Morphological Language Development

The Morphological Completion subtest on the Primary version consists of 38 cloze items designed to measure grammatic speaking or children's abilities to understand and produce grammatically correct utterances using common morphological forms. Children completed examiners' unfinished utterances using contextual information to select the proper morphological form of the target word. Responses were scored as either *correct* (1) or *incorrect* (0). Testing continued until children answered five consecutive items incorrectly. Internal consistency of the Primary version's Morphological Completion subtest at Time 1 was $\alpha = .68$ for the current sample.

The Morphological Comprehension subtest on the Intermediate version consists of 50 items and six "foil" items designed to measure grammatic listening or children's abilities to distinguish between grammatically correct versus incorrect utterances. Examiners read prompts to children and asked them to decide if the sentence they heard was correct or incorrect. In accordance with standard testing procedures for this subtest, the first 10 items (and "Foil" Items a and b) were administered to all children. Beginning with Item 11, a ceiling was reached if children answered any three items incorrectly in a group of five consecutive items. Also in accordance with standard testing procedures, testing was discontinued for any children who answered incorrectly to more than one "foil" item. For these children, the Morphological Completion subtest on the Primary version was then administered. Internal consistency of the Intermediate version's Morphological Comprehension subtest at Time 1 for the current sample was $\alpha = .92$.

Syntactic Language Development

The Syntactic Understanding subtest on the Primary version consists of 30 items in which examinees are given a verbal prompt and asked to select the picture that best matches the verbal prompt from a three-picture array. This subtest is designed to measure grammatic listening or children's abilities to understand the meaning of sentences

Table 3. Subtests from each version with number of items and administration rules.

Primary version subtest (items)	Intermediate version subtest (items)	Drop-back rule
Picture Vocabulary (34)	Picture Vocabulary (80)	Missed example item (i.e., child did not understand task)
Morphological Completion (38)	Morphological Comprehension (50)	Missed > 1 foil item (i.e., test not scorable and discontinued)
Syntactic Understanding (30)	Sentence Combining (30)	Missed first three items (i.e., immediate ceiling)

Note. Parentheses indicate the maximum number of items per subtest. The drop-back rule is the number of items the student would fail before being moved from the Intermediate task to the Primary task (left column). For the Intermediate version's Picture Vocabulary subtest, if the student did not understand the single example item, the student was then administered the Primary version.

using syntactic and morphological cues. After administering two sample items, testing began with Item 1 and continued until children answered five consecutive items incorrectly. Internal consistency of the Primary version's Syntactic Understanding subtest at Time 1 for the current study sample was $\alpha = .91$.

The Sentence Combining subtest on the Intermediate version consists of 30 items designed to measure grammatic speaking or children's abilities to understand and produce grammatically correct utterances using common morphological forms. Examiners read two or more sentences aloud and asked children to combine the sentences into a single, grammatically correct utterance that was as brief as possible. Responses were scored as either *correct* (1) or *incorrect* (0). In accordance with standard testing procedures for this subtest, after administering a sample item, testing began with Item 1 and was discontinued after examinees responded incorrectly to three consecutive items. For those children who reached an immediate ceiling on the Sentence Combining subtest (i.e., responded incorrectly to the first three items), the Syntactic Understanding subtest on the Primary version was then administered. Internal consistency of the Intermediate version's Sentence Combining subtest for the current sample at Time 1 was $\alpha = .91$.

Analysis

The analysis of the current article evaluates the structural validity of the TOLD-4 both within test version and across test versions over time. We used CFA to evaluate the structural validity and measurement equivalence of the tests—the extent to which each subtest measures a latent factor of language on a consistent metric for growth across versions and grades. In order to link the three subtests across versions and grades, student response data were divided into six administration groups based on which version they received in their particular grades (see Table 1).

Because the accelerated cohort design with "drop-back" testing was used in the current study, there was sufficient overlap to evaluate which measurement properties were consistent across the Primary and Intermediate versions of the TOLD-4. Specifically, there was overlap (a) within test version over time (e.g., both third and fourth graders received the Intermediate version), (b) across versions over time (e.g., some second graders dropped back from the Intermediate to Primary version, and some first graders progressed from the Primary to Intermediate version as they moved into second grade), and (c) across groups within grades (e.g., Group 5 included both third and fourth graders who received the Intermediate version).

The overlap in the design allows us to fit a CFA with the previously hypothesized properties of structural validity, cohort equivalence, and measurement equivalence. If these hypotheses are true, then the latent factor scores we develop are longitudinally consistent and reflect true differences across groups as well as individual change over time. Details for these assumptions are given in the Appendix, along with the sequence of models testing these assumptions.

These hypotheses require that statistical parameters of the model be held consistent across time and across groups within grades, so that children in different grades (or over time) may differ or grow. These consistent statistical parameters constitute *measurement equivalence*, a standard approach in CFA (Brown, 2015; Little, 2013; Rock, 1982; Vandenberg & Lance, 2000). All models were fit using Mplus Version 8 (Muthén & Muthén, 2017) with full information maximum likelihood estimation for missing data.

Because the model is large and complex, covering six groups at two time points each on six measures each, we present it in two figures. Figure 3 presents Administration Groups 1–4 for students in Grades 1–3. Group 1 received the Primary version in Grades 1 and 2; Group 2 received the Primary version in Grade 1 but the Intermediate version in Grade 2. Group 3 received the Primary version in Grade 2 but the Intermediate version in Grade 3. Group 4 received the Intermediate version in both Grades 2 and 3.

Figure 4 presents the same model, extended to Groups 5 and 6, who received the Intermediate version in Grades 3–4 and 4–5, respectively. The parameters for Group 5 in Grade 3 can be seen to match those for Groups 3 and 4 in Grade 3.

Within each group, there are the three subtests (rectangles) in two grades. The circles represent the general factor of language for that subtest on that version—Primary or Intermediate—in the first versus second versus third grade.

Each of the Greek letters corresponds to the statistical parameters of the model: λ is a factor loading, relating the test to the factor; ψ is a factor variance or covariance for the particular grade; α is the mean of the factor for that grade; v is the regression intercept for that test (model-predicted mean); and θ is a residual variance or covariance for test-specific error or the relation of error in that test over time. These parameters are held equal within test version and across time and groups to ensure that the factors are measured consistently (see Appendix).

Figures 3 and 4 show the measurement parameters to be equal over time within version across grades and groups (factor loadings and intercepts are consistent), and student parameters are consistent within grade across groups (factor mean and variances are equal) but are allowed to differ between grades for growth. More details on these assumptions and parameters can be found in the Appendix.

Results

Table 4 presents means and standard deviations for total scores on each of the six subtests at the two time points in the six groups. As suggested in Figure 1, there is general growth over time, but the overlap of versions in Grades 1 and 2, across Administration Groups 1–4, makes comparisons difficult (see Figure 1). Model fit in SEM is a large and complex issue, but we adopt the general guidelines of acceptable fit (comparative fit index [CFI] > .90, root mean square error of approximation [RMSEA]

Figure 3. Specification of the measurement model for the first four administration groups, Grades 1–3. Prim = Primary; Int = Intermediate; PV = Picture Vocabulary; MC = Morphological Completion, Primary/Morphological Comprehension, Intermediate; SU = Syntactic Understanding, Primary; SC = Syntactic Comprehension, Intermediate.

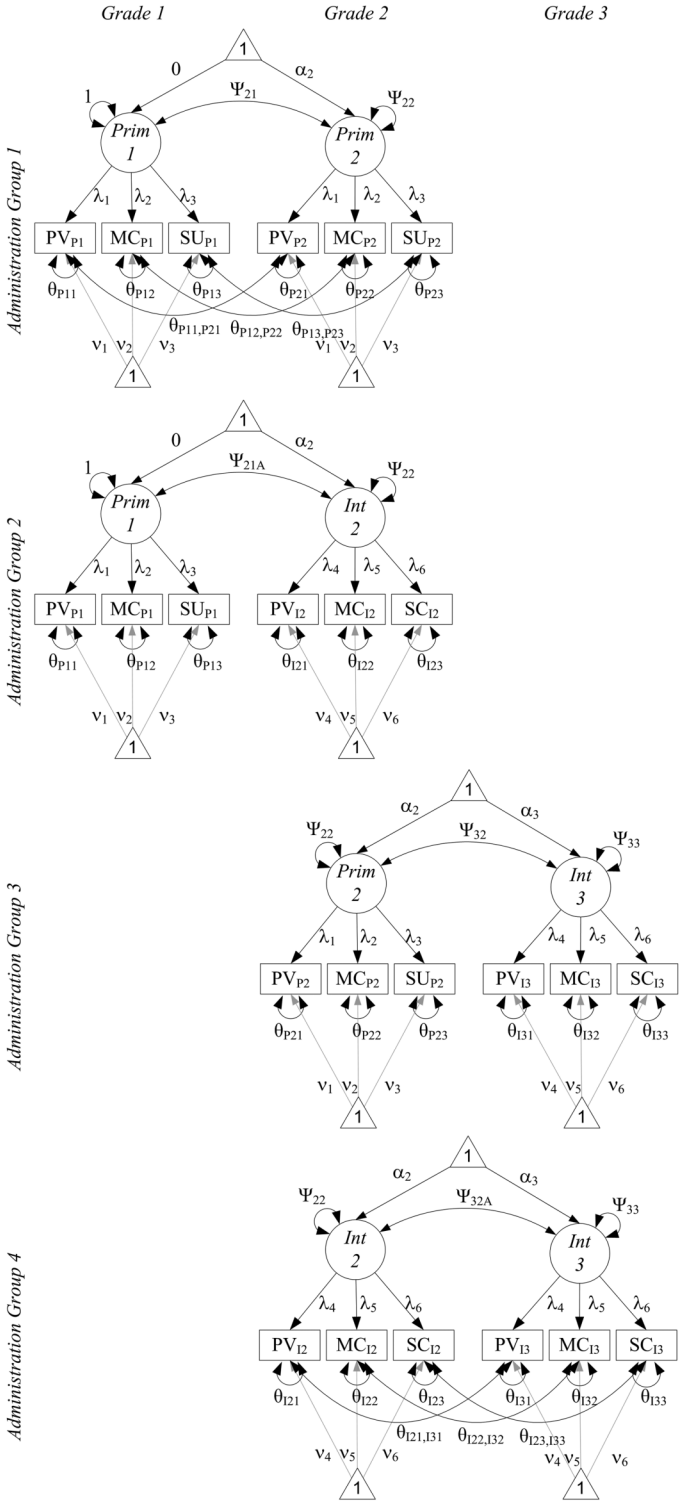
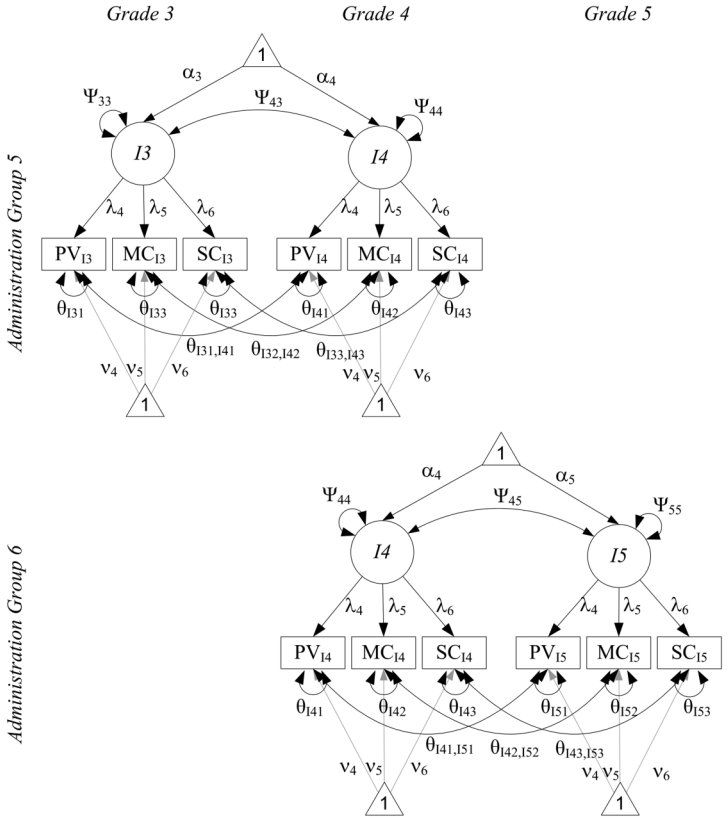


Figure 4. Specification of the measurement model for the final two administration groups, Grades 3–5. Each rectangle represents a test, followed by a subscript for test version (P = Primary; I = Intermediate) and grade designation (1–5). PV = Picture Vocabulary; MC = Morphological Completion, Primary/Morphological Comprehension, Intermediate; SU = Syntactic Understanding, Primary; SC = Syntactic Comprehension, Intermediate.



< 0.08) and excellent fit (CFI > .95, RMSEA < 0.05; see discussions by Little, 2013; Marsh, Hau, & Grayson, 2005; Marsh, Hau, & Wen, 2004). The full model, as shown in Figures 3–4, fits reasonably well, $\chi^2(109) = 189.8$, CFI = .95, RMSEA = 0.07. This model represents the final step in a sequence of tests—the most consistent and parsimonious—but details on tests of equality are given in the Appendix.

Table 5 shows the measurement parameters, loadings, and intercepts, which correspond to those in Figures 3 and 4.

Table 4. Descriptive statistics by administration group for Test of Language Development–Fourth Edition subtests.

			M			*SD*		
Group	**Grade**	**Version**	**Morph**	**Vocab**	**Syntax**	**Morph**	**Vocab**	**Syntax**
1	1	Primary	17.4	19.1	21.8	7.9	4.5	3.9
	2	Primary	21.4	21.1	24.1	7.4	4.9	2.9
2	1	Primary	14.6	18.2	21.3	7.3	4.4	3.7
	2	Intermediate	9.5	26.8	7.5	6.5	10.2	5.3
3	2	Primary	19.6	20.1	23.5	8.3	4.4	3.1
	3	Intermediate	13.9	33.1	10.2	9.9	11.3	6.2
4	2	Intermediate	9.7	28.2	6.7	6.3	9.5	5.7
	3	Intermediate	13.7	36.1	9.3	8.8	12.9	6.2
5	3	Intermediate	10.6	30.3	7.7	7.3	9.8	5.1
	4	Intermediate	12.8	37.5	10.3	8.5	9.4	5.8
6	4	Intermediate	13.2	35.8	10.0	8.3	9.4	5.8
	5	Intermediate	14.2	40.5	13.0	9.6	10.7	6.9

Note. Scores are total items correct (see Table 3). Correlation matrices for each group are available from the first author. Morph = Morphological Completion or Comprehension; Vocab = Picture Vocabulary; Syntax = Syntactic Understanding or Comprehension.

Table 5. Measurement parameters from the full scaling model.

Form	Subtest	Loading	(*SE*)	Intercept	(*SE*)
Primary	Morphological Completion	6.13	(0.40)	15.32	(0.42)
	Picture Vocabulary	3.17	(0.22)	18.20	(0.25)
	Syntactic Understanding	2.25	(0.17)	21.75	(0.19)
Intermediate	Morphological Comprehension	4.24	(0.50)	5.33	(0.70)
	Picture Vocabulary	7.51	(0.86)	21.48	(1.16)
	Syntactic Comprehension	3.12	(0.37)	4.20	(0.54)

Note. Standard errors are in parentheses. Model fit: $\chi^2(53) = 189.8$, comparative fit index = .953, root mean square error of approximation = 0.072, 90% CI [0.054, 0.088]. See Figures 3–4.

These parameters represent the regression of each test upon the latent factor of language and were held constant for the six subtests across grades and administration groups (i.e., six loadings in total).

The CFA model allows us to estimate a developmental scaled score, which is represented by a latent factor: Each student gets a factor score. Table 6 shows the factor means, variances, and correlations across student scaled scores in each grade. These were estimated relative to students in Grade 1 ($M = 0$, $SD = 1$). The first column shows steadily increasing means, at about 0.6 *SD* per year, with a slight slowing in Grade 5 (a gain of 0.4 *SD* from prior year), suggesting that, after controlling for subtest and version, we see relatively consistent growth over time. The variances are relatively consistent, close to 1.0 in each grade, suggesting that student growth trajectories in language do not exhibit strong convergence or fan-spread (like a Matthew effect). The correlations describe the longitudinal relations between the first year and the second year each student was measured. These were extremely high ($r = .84$–1.00), suggesting high stability of student ranking after controlling for measurement error in the separate tests.

Table 7 shows the standardized factor loadings (validity coefficients), residual variance (θ), within-test longitudinal lag correlation for the groups (where relevant), and R^2 values (model-based reliability). Standardized factor loadings are validity coefficients, reflecting the correlation between the variable and its intended factor. Residual variance is a measure of unexplained error, from which we can compute R^2. The within-test lag correlation represents the extent to which error is related over time within subtest—if these values are large, then the factor structure we have tested may be missing large, systematic sources of variance over time (i.e., there may be something other than the factor that explains relations across grades).

Table 6. Latent variances, correlations, and means.

Grade	*M*	(*SE*)	Variance	(*SE*)	Correlation
1	0.00	—	1.00	—	—
2	0.70	(0.08)	0.96	(0.13)	.89
3	1.33	(0.14)	1.35	(0.32)	.98
4	1.95	(0.19)	1.18	(0.28)	.84
5	2.36	(0.24)	1.36	(0.38)	1.00

Note. "Correlation" refers to the correlation between the current grade and the previous grade (see Figures 3–4). Dashes indicate an estimate is not relevant because it is fixed to identify the model. Standard errors are in parentheses.

The standardized factor loadings ($Mdn = 0.68$) suggest good validity for these tests as indicators of a general factor of language. The lag correlations were low to moderate, suggesting that the factor model adequately captured longitudinal change, as opposed to test-specific error that may have been consistent over time. The R^2 values show a range of reliability (.26–.71), with overall moderate reliability ($M = .47$).

Figure 5 presents the resulting developmental scaled scores of this equating model. The scaled scores for each student in this model were calculated in Mplus (FSCORES output option) and then graphed. The vertical axis shows the value of the language factor score over grades (horizontal axis). Each student is represented by a gray line. The dark line is a moving average (loess) line for these data, which corresponds closely to the factor means shown in Table 6. The overall trend for these lines is a steady, almost linear increase in language across grades, which stands in stark contrast to the inconsistencies in the nonequated, observed scores shown earlier in Figures 1–2.

Discussion

Why are standard scores inadequate for longitudinal models? Standard scores fail to have appropriate longitudinal properties because they are standardized relative to a fixed age group (i.e., 8 vs. 9 years), not to a developmental metric of growing a certain number of units per year. That is, the "ruler" for standardized scores is slightly different: It is relative to the standard deviation of the particular collection or cohort of age, regardless of whether those children participated in multiple time points or how much they might have grown. Such limitations are widely known in reading research (Seltzer et al., 1994) and in developmental psychology (Horn & McArdle, 1992; Meredith & Horn, 2001). By modeling consistent measurement properties in a longitudinal sample, we are able to index individual change on an appropriate developmental metric.

For valid and ethical measurement, children of different ages are often given tests of differing difficulty and

Table 7. Group-specific parameters.

Group	Subtest	Standardized factor loading		Residual variance		Lag correlation	R^2
1	MC1	0.81	(0.03)	20.22	(3.28)	.41	.65
	PV1	0.70	(0.04)	10.23	(1.21)	.24	.50
	SU1	0.60	(0.04)	8.78	(0.93)	–.08	.37
	MC2	0.78	(0.03)	23.18	(3.40)	—	.61
	PV2	0.72	(0.04)	8.90	(1.17)	—	.52
	SC2	0.72	(0.04)	4.56	(0.58)	—	.52
2	MC1	0.81	(0.03)	20.22	(3.28)	—	.65
	PV1	0.70	(0.04)	10.23	(1.21)	—	.50
	SU1	0.60	(0.04)	8.78	(0.93)	—	.37
	MC2	0.65	(0.06)	23.71	(6.46)	—	.42
	PV2	0.83	(0.07)	25.44	(10.12)	—	.68
	SC2	0.51	(0.06)	27.38	(7.01)	—	.26
3	MC1	0.78	(0.03)	23.18	(3.40)	—	.61
	PV1	0.72	(0.04)	8.90	(1.17)	—	.52
	SU1	0.72	(0.04)	4.56	(0.58)	—	.52
	MC2	0.56	(0.03)	52.77	(4.69)	—	.32
	PV2	0.79	(0.03)	44.98	(6.55)	—	.63
	SC2	0.58	(0.04)	26.03	(2.44)	—	.34
4	MC1	0.58	(0.04)	34.09	(2.83)	.52	.34
	PV1	0.80	(0.04)	30.83	(5.22)	.47	.64
	SU1	0.56	(0.04)	20.29	(1.64)	.36	.32
	MC2	0.56	(0.03)	52.77	(4.69)	—	.32
	PV2	0.79	(0.03)	44.98	(6.55)	—	.63
	SC2	0.58	(0.04)	26.03	(2.44)	—	.34
5	MC1	0.65	(0.03)	34.09	(2.83)	.33	.42
	PV1	0.84	(0.03)	30.83	(5.22)	.23	.71
	SU1	0.63	(0.03)	20.29	(1.64)	.39	.39
	MC2	0.54	(0.04)	52.77	(4.69)	—	.29
	PV2	0.77	(0.03)	44.98	(6.55)	—	.60
	SC2	0.55	(0.04)	26.03	(2.44)	—	.31
6	MC1	0.62	(0.04)	34.09	(2.83)	.56	.38
	PV1	0.83	(0.03)	30.83	(5.22)	.31	.68
	SU1	0.60	(0.04)	20.29	(1.64)	.24	.36
	MC2	0.56	(0.05)	52.77	(4.69)	—	.32
	PV2	0.79	(0.04)	44.98	(6.55)	—	.63
	SC2	0.58	(0.05)	26.03	(2.44)	—	.34

Note. Subtest abbreviations (see Table 3): MC1 = Morphological Completion, Primary; PV1 = Picture Vocabulary, Primary; SU1 = Syntactic Understanding, Primary; MC2 = Morphological Comprehension, Intermediate; PV2 = Picture Vocabulary, Intermediate; SC2 = Sentence Combining, Intermediate. Em dashes indicate that an estimate is not relevant. "Lag correlation" refers to the within-group error correlation over time for students who took the same test version across 2 years (paired for the three subtests). Such an error correlation was not modeled for Administration Groups 2 and 3 because they took two different versions of the test in subsequent years (Figures 3–4).

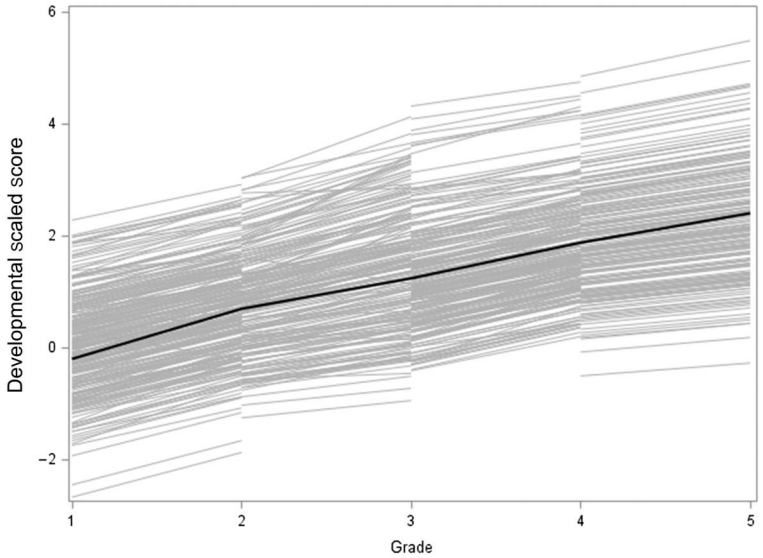
composition. However, to understand longitudinal development, tests must have a consistent metric over time, which means psychometric equating studies, if not done by the publisher in test development, need to be undertaken by researchers in the field. Psychometric equating studies allow for the creation of developmental scaled scores, without which studies are either piecemeal or highly problematic for longitudinal conclusions. Modeling language development longitudinally with a consistent metric for growth allows for not only the description of children's language development but also the exploration of relations between language development and many other outcomes of interest.

Conclusions

The current study involved two versions of a language test, the TOLD-P:4 and the TOLD-I:4, each with three subtests. Without a longitudinally consistent scaled score, these six subtests would be essentially incommensurable, making conclusions about growth difficult, if not impossible. Using a cohort sequential design across five grades, the overlapping administrations of these six subtests allowed us to fit a latent variable model (longitudinal CFA) that jointly scaled the tests, across versions, grades, and cohorts to yield a developmental scaled score for language. The model fits well, with good parameter estimates, suggesting that our assumptions were reasonable about the joint functioning of these subtests. Whereas the observed total and standard scores from the subtests suggested inconsistent and problematic patterns of growth (see Figures 1 and 2), the scaled scores from this model suggest a continuous average pattern of language growth across Grades 1–5 (see Figure 5).

Results from such scaled scores allow longitudinal analyses across all five grades. For example, we have examined the longitudinal relations of language with reading and dialect (Washington et al., 2018) and the extent to which there may be gender differences in language growth (Washington et al., in press). Without a longitudinal scaled

Figure 5. Resulting developmental scaled scores (student factor scores) from the equated model. The gray lines represent individual students (n = 865). The dark line represents a moving average (loess). See Table 6 for model-estimated means and variances.



score for language, examinations of these sorts of questions would likely end up being piecemeal and incomplete (e.g., several cross-sectional analyses), rather than the full, 5-year growth model analyses possible in such a design.

These results in prior articles, however, were only a cursory application of an ideal model. The current article takes a rigorous, step-by-step examination of the assumptions of that model for language across tests. The previous model essentially took the structure of the final, equated model and used it to estimate student scores. The current examination details each of the previously presumed steps and lays out their conceptual implications in order to didactically illustrate how we can (and should) use multiple measures of constructs in order to develop longitudinally consistent metrics. The current results suggest that those previously used scores are valid.

Although the technical details of the statistical model for longitudinal and across-group equating are complex, the motivation is simple: We wish our tests to indicate their intended construct in an equivalent manner over time, across groups, and across versions of the test. In the current study, we fit a confirmatory model testing these wishes as research hypotheses, imposed as statistical constraints on the test scores. The results suggest that these six subtests of the TOLD-P:4 and TOLD-I:4 reasonably fit such a model of equivalence. In particular, this model implies that for students similar to the current sample:

1. All of the six subtests jointly measure general language proficiency.
2. When scaled by this model, these tests measure language in a longitudinally consistent manner, across grades and versions.
3. Growth in language is relatively consistent on average, about half of 1 *SD* per year (a gain of 0.4–0.6 *SD* per year).

It is worth noting that the model imposes measurement equality across groups and versions but imposes no constraints on the shape of growth (e.g., it is not forced to be linear). The model is, in effect, a latent variable version of repeated measures, with two time points per child, linked across cohorts from Grades 1 to 5. The result in Figure 5 shows that, when we assume measurement equivalence for these tests, growth in language appears reasonably linear in Grades 1–5.

Limitations

Population Generalizability

It is important to note some limitations of the current study. The students in the current sample were nearly all African American, and so the equating across versions and over time may be different for other cultural and linguistic groups (although the same measurement principles apply). It will be informative to test the TOLD-4 in other populations to ensure its construct validity, and longitudinal measurement invariance holds across the different versions of the test.

These sorts of individual differences between students can be extended to clusters of students and should also be examined in future research. For example, the current study ignored classroom variation. Multilevel models examining variation across classrooms or differences between classrooms and schools could be highly informative. It should be noted that the current design involved 154 different teachers,

with students moving into 334 unique, 2-year classroom combinations. Because this was a student-based longitudinal study, classmates in students' second year were not measured, so a full, cross-classified model would imprecisely estimate clustering effects in the second year. Furthermore, there are no simple classroom corrections applicable to such cross-classified models. Based on previous multilevel work, it is unclear what effect ignoring classroom clustering might have. We suspect that classroom-level relations are higher and more consistent for language and literacy outcomes (Branum-Martin et al., 2006; Mehta, Foorman, Branum-Martin, & Taylor, 2005; Mehta & Neale, 2005), potentially making the current examination fit better than might be found in a cross-classified model of students switching classrooms. For our present purpose, we wished to demonstrate a psychometric equating model for a developmental score. Classroom structure in a longitudinal design is left to further work, where adequate sampling of classmates in follow-up years might be helpful.

Similarly, clusters of individuals may demonstrate meaningful differences over time (i.e., cohorts may differ from each other). Although the current model fits well and suggests cohort differences are negligible, a design with more time points would be able to more rigorously isolate cohort differences from longitudinal change (Meredith & Tisak, 1990; Schaie & Baltes, 1975).

Considerations for Measuring Language

The current study used only three tests per time point, with no other standardized measures of language. Each factor was measured by only three tests (a just-identified CFA model), using only selected features of semantics and grammar. Moreover, the TOLD-4 does not specify drop-back rules between tests. It is possible that other subtests (e.g., Relational Vocabulary or Word Ordering) or tasks from other tests could have different results—however, there is reason to expect that a general factor of language is reasonable (Foorman, Koon, Petscher, Mitchell, & Truckenmiller, 2015). Similar models with more tests—and modeled at the item level—may be more informative to evaluate the extent to which a general factor of language can be measured in a longitudinally consistent manner.

Similarly, the current examination is at the level of the test score, not at the level of items (Embretson, 2006, 2007; Holland & Dorans, 2006). It is possible that some items may function differently at some grade levels (i.e., exhibit differential item functioning). Given that item-level models would be nested within the current score-level equating model, it is unlikely that such differential functioning would be large in the current sample (i.e., the current models are good enough that it is unlikely that item-level models would uncover large sources of bias or misfit). Differences in test functioning might also exist for subgroups of examinees (e.g., across genders: the current design would be doubled and could be tested for invariance across genders). Such complexities, although important for decisions about test design, are left to future research, with the current framework providing guidance for how to examine those.

The current examination measured students only once per year and only near the end of the year. A design with more time points per student may achieve better description of the shape of growth in language or individual variation in trajectories. Longitudinal models can accommodate person-specific times of observation (Mehta & West, 2000), and modeling time correctly at the individual level, rather than in fixed waves, can help to better describe the nature of growth and disability (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996). The current approach could either be extended to individual times of administration, or more simply, these equated scores could be used to better understand individual trajectories in language over elementary grades (Washington et al., 2018, in press).

Considerations for Equating

The current study was not designed as an equating study with random assignment to test versions within appropriately overlapping ages. Instead, children were administered versions based on their age, and those above 8 years old were administered the Primary version if they failed to perform adequately on the Intermediate version. The current analysis only had overlap across forms in two groups involving second grade (Groups 2–3, $n = 280$), and only a few students received both forms ($n = 104$). Our calculations (MacCallum, Browne, & Sugawara, 1996) suggest we have excellent power to distinguish even mediocre models from good ones (i.e., RMSEA = 0.08 from RMSEA = 0.05). Although assignment to versions was based on age and performance, we would not expect bias in the measurement parameters—such assignment should get modeled as differences in the estimated latent ability. An ideal design for such questions would randomly assign test versions to students near the age cutoff (e.g., all 7- to 8-year-old students take both versions or take some overlapping subsample of both versions). As it stands, the fit and properties of the current model do not indicate problems with bias or nonrandom assignment. The overlap across cohorts and test versions was likely practical and appropriate (see Table 1). However, a model with more indicators per factor and with greater overlap—randomly assigned—may provide a more rigorous test of the equality of language tests across forms.

There are a number of ways to equate across tests. The current approach using multiple-group CFA assumes linear relations among tests and a multivariate normal distribution but is rigorous in its statistical implications (Horn & McArdle, 1992; Rock, 1982). Nonparametric equating methods (e.g., equipercentile) might have less strict statistical assumptions but may depend on assuming equivalence of sampling—that groups of students are exchangeable (Kolen, 2006; Petersen et al., 1989). Equipercentile equating has been used to adjust for version differences in reading fluency (Francis et al., 2008).

Recommendations for Practice

Multiple-group, longitudinal CFA is a complex statistical method, and large longitudinal data are difficult and expensive to collect, even if from an accelerated design. Therefore, for applied researchers, clinicians, and teachers, we offer the following recommendations. These recommendations are ordered from easiest to most difficult to implement.

Choose Tests That Have Developmental, Longitudinal Scales

Some publishers may have used item response theory to create a developmental scaled score (e.g., the *W* score on Woodcock-Johnson tests). The technical manual for the test or the online scoring program will provide the conversion from the total score to the scaled score, which would be appropriate for indexing longitudinal student gains. Teachers, clinicians, and researchers interested in longitudinal comparisons will be well served to choose a test with a developmental scaled score available.

Use Overlapping Tests or a Single Measure That Is Consistent

Some tests may represent methods or content that is important to the research question but may not be longitudinally linked. If some tests differ by age, using another test that has a developmental scaled score can provide an anchor. This anchor test can be used to evaluate change in longitudinally inconsistent (but substantively interesting) tests. Such a design would require statistical linking, similar to that used in the current study.

Do an Equating Study

If it is known that there is a crucial test that is not longitudinally consistent, then the study can be designed to make equating possible, with the appropriate overlap in versions and groups, or an anchor test (see above). There are many techniques for equating (Holland & Dorans, 2006; Kolen & Brennan, 2014), including the current longitudinal CFA approach. The goal of an equating study is to ensure that enough students get enough overlapping versions or items so that dependable equating can be done (i.e., in a statistical model or in a linking paradigm).

Redesign the Test or Build Your Own

If the construct is crucial to your study but the older and younger versions of existing tests are not equated, a joint test form with mixed items could potentially be designed. In designing such a measure, careful attention should be given to ensure that items are developmentally appropriate and that item formats are practical and understandable to examinees. Designing items, pilot testing, and validating item responses is an expensive, time-consuming process, but such investment is necessary to develop tests with appropriate reliability and validity for longitudinal research.

Acknowledgments

The research reported here was supported by the National Institute of Child Health and Human Development, through Grant 1R24HD075454 (Julie Washington, PI).

References

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Branum-Martin, L., Mehta, P. D., Fletcher, J. M., Carlson, C. D., Ortiz, A., Carlo, M. S., & Francis, D. J. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology, 98*(1), 170–181. https://doi.org/10.1037/0022-0663.98.1.170

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: Guilford Publications.

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Embretson, S. E. (2006). The continued search for nonarbitrary metrics in psychology. *American Psychologist, 61*(1), 50–55. https://doi.org/10.1037/0003-066X.61.1.50

Embretson, S. E. (2007). Impact of measurement scale in modeling development processes and ecological factors. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 63–87). Mahwah, NJ: Erlbaum.

Foorman, B. R., Koon, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. *Journal of Educational Psychology, 107*(3), 884–899. https://doi.org/10.1037/edu0000026

Francis, D. J., Santi, K. L., Barr, C., Fletcher, J. M., Varisco, A., & Foorman, B. R. (2008). Form effects on the estimation of students' oral reading fluency using DIBELS. *Journal of School Psychology, 46*(3), 315–342.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*(1), 3–17.

Hammill, D. D., & Newcomer, P. L. (2008a). *Test of Language Development–Primary: Fourth Edition*. Austin, TX: Pro-Ed.

Hammill, D. D., & Newcomer, P. L. (2008b). *Test of Language Development–Intermediate: Fourth Edition*. Austin, TX: Pro-Ed.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: National Council on Measurement in Education & American Council on Education.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*(3), 117–144.

Jöreskog, K. G. (1970). *Simultaneous factor analysis in several populations*. Princeton, NJ: Educational Testing Service. Retrieved from http://eric.ed.gov/?id=ED055110

Jöreskog, K. G. (1979). Simultaneous factor analysis in several populations. In J. Magidson (Ed.), *Advances in factor analysis and structural equation models* (pp. 189–206). Cambridge, MA: Abt Books.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test–Second Edition (KBIT-2)*. Bloomington, MN: Pearson.

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York, NY: Guilford.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: National Council on Measurement in Education & American Council on Education.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating, scaling, and linking: Methods and practices*. New York, NY: Springer.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York, NY: Springer.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY: Guilford Press.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130–149.

Marsh, H. W., Hau, K.-T., & Grayson, D. (2005). Goodness of fit in structural equation models. In A. Maydeu-Olivares & J. J. McArdle (Eds.), *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Mahwah, NJ: Erlbaum.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(3), 320–341. https://doi.org/10.1207/s15328007sem1103_2

McArdle, J. J., & Grimm, K. J. (2010). Five steps in latent curve and latent change score modeling with longitudinal data. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Longitudinal research with latent variables* (pp. 245–273). Berlin, Germany: Springer.

McArdle, J. J., & Hamagami, F. (1991). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 276–304). Washington, DC: American Psychological Association.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Mehta, P. D., Foorman, B. R., Branum-Martin, L., & Taylor, W. P. (2005). Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. *Scientific Studies of Reading, 9*(2), 85–116.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10*(3), 259–284.

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods, 5*(1), 23–43.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*(4), 525–543.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203–240). Washington, DC: American Psychological Association.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*(1), 107–122. https://doi.org/10.1007/bf02294746

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York, NY: Macmillan.

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (Vol. 154, pp. 163–180). Newbury Park, CA: Sage.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163. https://doi.org/10.2307/271063

Rock, D. A. (1982). Equating using the confirmatory factor analysis model. In P. W. Holland & D. B. Rubin (Eds.), *Test equating* (pp. 247–257). New York, NY: Academic Press.

Schaie, W., & Baltes, P. B. (1975). On sequential strategies in developmental research. *Human Development, 18*(5), 384–390.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to choice of metric. *Educational Evaluation and Policy Analysis, 16*(1), 41–49.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70.

Washington, J. A., Branum-Martin, L., Lee-James, R., & Sun, C. (in press). Reading and language performance of low-income, African American boys in grades 1 to 5. *Reading & Writing Quarterly*. https://doi.org/10.1080/10573569.2018.1535777

Washington, J. A., Branum-Martin, L., Sun, C., & Lee-James, R. (2018). The impact of dialect density on the growth of language and reading in African American children. *Language, Speech, and Hearing Services in Schools, 49*(2), 232–247. https://doi.org/10.1044/2018_LSHSS-17-0063

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8*(1), 16–37. https://doi.org/10.1037/1082-989X.8.1.16

Appendix (p. 1 of 2)

Technical Appendix

The model is a longitudinal, multiple-group confirmatory factor analysis. Each group of students was measured up to two times, and grade levels overlap. This is a straightforward application of multiple-group SEM (Brown, 2015; Jöreskog, 1970, 1979; Vandenberg & Lance, 2000) for longitudinal data (Little, 2013). Students in cohorts measured only once, or who were absent or moved away, are assumed missing at random so that measures they did complete are used at the time points where they are present (McArdle & Hamagami, 1991).

Tests are evaluated for their equivalence at each time point and across groups, while students are assumed to be similar within grades but different over time. These are the standard, desired properties of measurement equivalence (Meredith, 1993; Vandenberg & Lance, 2000) and are detailed below in nontechnical language.

Structural assumptions of the model include the following (see Figures 3–4):

1. Each test measures a single factor: Although each test measures different aspects of language, they all measure a single factor of general language proficiency. Although the current analysis uses sum scores, this assumption applies all the way down to the individual items that comprise each test (McDonald, 1999).
2. Measurement is on a consistent metric: Pattern coefficients (factor loading: λ) are equal for a given subtest, in any grade (i.e., over time and across groups). There are only six factor loadings, one for each of three subtests across two versions.
3. Each test has a consistent zero point (regression intercept: ν) so that growth or mean difference between ages or groups can be indexed: Regression intercepts are equal for a given subtest version, in any grade (i.e., no mean bias). There are only six regression intercepts, one for each of three subtests across two versions.
4. Subtests have the same amount of error across groups within grades (residual variance: θ): Error does not differ within subtests across cohorts. There are 18 residual variances.
5. Subtest errors may be related over time: Residual covariance for the same child retaking the same subtest a year later may be allowed (Little, 2013). There are 12 residual covariances.
6. Students within a grade are exchangeable, as if drawn from a single population (cohort differences are ignorable):
   a. Latent means (α) are equal within grade, across administration groups (one per grade, 2–5).
   b. Latent variance (ψ) is equal within grade, across administration groups (one per grade, 2–5).

The first four of these assumptions are standard for measurement equivalence in multiple-group or longitudinal factor analysis, respectively corresponding to metric, scalar, and residual invariances (Meredith, 1993; Vandenberg & Lance, 2000). The fifth assumption is a standard for equivalent groups (i.e., second graders are a single population), as is common in equivalent groups equating (Holland & Dorans, 2006). The technical details of specifying the parameters for these assumptions may be found in texts that cover multiple-group SEM (Brown, 2015; Kline, 2016; Little, 2013).

Invariance testing. Table A1 presents the full sequence of model tests for longitudinal and across-group equivalence (Brown, 2015; Little, 2013; Vandenberg & Lance, 2000). The first model is a baseline null model against which the others are compared (Little, 2013). Each of the subsequent models is a restriction of the model that comes before it, with the configural testing the single factor of general language, metric testing the equality of factor loadings, scalar testing the equality of intercepts, and uniqueness testing the equality of residual variances. All of these models fit reasonably on their own (CFI > .95, RMSEA < 0.08).

The sequence of models fails the chi-square test of equality at each step, but this test is known to be excessively conservative, especially for so many groups in a large sample (Chen, 2007). Moreover, there are no adapted guidelines for judging fit (e.g., change in CFI) for so many groups (Chen, 2007). We present these results for the sake of transparency, in order to inform a discussion of statistical testing versus practical measurement. It is worthwhile to note that, in Table A1, each model drops in BIC more than 100 units, indicating large increases in fit, relative to parsimony (Raftery, 1993, 1995). Given the overall good fit of each model and the large changes in BIC, we suggest that the fully restricted uniqueness model is a reasonable statistical model that matches our a priori theory and the design of the project.

Sensitivity analyses. Table A2 reports the fit of two models we used as sensitivity checks, one for the problems in the standard model (see Table A1) and the other for the mixture of test versions by students. The "Note" column in Table A1 shows that there were some estimation problems in each of the models. The longitudinal correlation between the latent factors estimated at greater than unity in Group 5 for all models. One of the tests had a negative residual variance in the configural model, and the latent correlation for Group 6 estimated at greater than unity—however, these two problems disappeared under further restrictions and were not present in the final uniqueness model.

Table A1. Model fit

Model	Parameters	*df*	χ^2	BIC	CFI	TLI	IFI	RMSEA	[90% CI]	Note
Null	21	141	1846.4	25,794	0	0	0	0.290	[0.278, 0.302]	
Configural	126	36	39.9	24,698	.998	.991	.998	0.027	[0.000, 0.067]	PV2, G5
Metric	100	62	89.9	24,572	.984	.963	.984	0.056	[0.027, 0.080]	G5, G6
Scalar	74	88	152.8	24,459	.962	.939	.963	0.071	[0.052, 0.090]	G5, G6
Uniqueness	53	109	189.8	24,354	.953	.939	.954	0.072	[0.054, 0.088]	G5

Note. The null model is a model of independence for repeated-measures data, which is more appropriate than the standard null in SEM software (Little, 2013; Widaman & Thompson, 2003). BIC = Bayesian information criterion; CFI = comparative fit index; TLI = Tucker–Lewis index; IFI = incremental fit index; RMSEA = root mean square error of approximation, with 90% confidence interval (CI); PV2 = Picture Vocabulary in Group 2, Grade 2, had a mildly negative residual variance; G5 = latent correlation greater than unity in Group 5; G6 = latent correlation greater than unity in Group 6.

In order to test the validity of our model in the face of these problems, we fit the uniqueness model only to Groups 1–4, in order to avoid the problematic groups. Table A2 shows that this model fit very well. Additionally, we fit models (not reported here) in which we forced the relation between the latent factors in Group 5 to be perfect, and this model had similarly good fit. We are therefore reasonably confident about the stability of our results.

Because the sample had students who took a mixture of test versions (n = 76; see Table 2), we also fit the full uniqueness model using only students who did not receive multiple versions (n = 789). This model on "pure" students fit similarly well (see Table A2), suggesting that our results are not merely an artifact of strange overlapping of tests.

Table A2. Sensitivity: fit for alternative models.

Model	Parameters	*df*	χ^2	CFI	TLI	IFI	RMSEA	[90% CI]	Note
Uniqueness, Groups 1–4 only	41	67	90.3	.974	.966	.987	0.055	[0.018, 0.082]	
Uniqueness, pure students only	53	109	175.8	.958	.946	.962	0.068	[0.049, 0.086]	G5, G6

Note. Both models are tested against their respective longitudinal null model, not shown here (Little, 2013; Widaman & Thompson, 2003). The "pure students" model was fit to test the model for reasonableness, excluding students with mixed administrations within groups. The "Groups 1–4" model was fit to test the model without excessively high latent longitudinal correlations in Groups 5–6. CFI = comparative fit index; TLI = Tucker–Lewis index; IFI = incremental fit index; RMSEA = root mean square error of approximation; CI = confidence interval; G5 = latent correlation greater than unity in Group 5; G6 = latent correlation greater than unity in Group 6.