Article

LANGUAGE
TESTING

Language Testing
28(3) 383–400

DOI: 10.1177/0265532211404190
ltj.sagepub.com
SAGE

Issues in vertical scaling of a K-12 English language proficiency test

Dorry M. Kenyon, David MacGregor, and Dongyang Li
Center for Applied Linguistics, USA

H. Gary Cook
Wisconsin Center for Educational Research, USA

Abstract

One of the mandates of the No Child Left Behind Act is that states show adequate yearly progress in their English language learners' (ELLs) acquisition of English language proficiency. States are required to assess ELLs' English language proficiency annually in four language domains (listening, reading, writing, and speaking) to measure their progress; they are also required to report on a composite comprehension measure. Often the clearest way to effectively monitor students' progress is to measure assessment results across grades on the same scale. In measurement terms, scores from tests across all grade levels can be put on the same scale using vertical scaling. In addition, to help stakeholders understand and interpret the results, these scale scores are often interpreted in terms of proficiency levels. In this article, we use the vertical scaling of WIDA ACCESS for ELLs®, a large-scale K-12 Academic English Language Proficiency assessment to illustrate measurement and practical issues involved in this technique. We first give background on the need for vertical scaling. We then assess the literature on vertical scaling and describe the procedures used for WIDA ACCESS for ELLs® to vertically scale test scores and interpret the results in terms of the WIDA ACCESS for ELLs® Proficiency Scale. Next we review several studies that have been conducted to gauge the effectiveness of that scaling. We end the paper with a discussion of the broad issues that arise from vertical scaling.

Keywords

English language proficiency tests, test development, test linking, vertical scaling

Corresponding author:
Dorry Kenyon, Center for Applied Linguistics, 4646 40th Street, NW, Washington, DC 20016, USA.
Email: dkenyon@cal.org

Background

The No Child Left Behind (NCLB) Act (No Child Left Behind Act of 2001; see also Bunch, this volume, and Bailey & Huang, this volume) mandates that all English language learners (ELLs) in grades K-12 be tested in their proficiency in four domains of English language: speaking, writing, listening and reading. Results from these tests are used to establish two of three types of Annual Measurable Achievement Outcomes (AMAOs), which are reported by states to the federal government:

> (i) at a minimum, annual increases in the number or percentage of children making progress in learning English;
>
> (ii) at a minimum, annual increases in the number or percentage of children attaining English proficiency by the end of each school year, as determined by a valid and reliable assessment of English proficiency; and
>
> (iii) making adequate yearly progress in academic achievement tests for limited English proficient children. (www.ed.gov/policy/elsec/leg/esea02/pg42.html; accessed January 5, 2009)

In addition to these federal NCLB requirements that are mandatory for states, districts and schools can use results from English language proficiency tests to assist them in making decisions on the placement of individual students into ELL services, in the evaluation of the effectiveness of their programs in helping their ELLs make progress in acquiring English, and in the determination of when to exit individual students from ELL services. Since results from these English language proficiency assessments are used to make decisions about students across many years, it is most useful to states, districts, and schools if the scores on the English language proficiency assessment are on the same score scale from year to year, as students progress anywhere along the continuum from Kindergarten to grade 12. In other words, in order to be able to compare scores, performances on the test will be most useful if they are reported on the same score scale when, for example, the student is in Kindergarten, first, second, and third grades, or in seventh, eighth, and ninth grades.

Clearly the same test forms cannot be administered across a continuum spanning grades K to 12. Instead, age- and developmentally-appropriate tests must be designed for each grade or clusters of grades. Doing so, however, requires that results on the forms be *vertically* scaled. That is, raw scores on the test forms at the different grade levels must be converted to a single underlying scale. Unlike *horizontal* scaling, in which different test forms (or versions) of a test are equated for a common population of students within a single grade, vertical scaling presents the challenge of linking tests, or putting scores on the same scale, across populations that developmentally can be very diverse. Vertical scaling, then, poses some technical issues and challenges, in particular because, as Young (2006) writes in a summary chapter on issues in creating vertical scales, ‘there is no consensus on what constitutes best professional practice for developing and validating vertical scales’ (p. 469).

The goal of this paper is to present an example of creating a vertical scale for a large-scale English language proficiency test. We first discuss various options available for creating a vertical scale based on a review of literature. We then describe the method used for creating the vertical scale for the WIDA Consortium's ACCESS for ELLs®, with a special focus on the issues particular to language testing, including the difficulty of vertically scaling performance-based tests in writing and speaking. We then briefly evaluate the results of this linking based on several years of operational data from forms that are horizontally equated to the previous year's form. Finally, we discuss some of the remaining issues that arise from vertical scaling.

Vertical scaling literature review

In educational achievement and aptitude assessments, it is common practice to use multiple test forms with different degrees of difficulty so that each form targets a particular group (classified through grade, age, or proficiency level) of examinees. To make scores comparable across these forms, a common scale score is usually established for the series of test forms, in particular so that growth can be measured. Kolen and Brennan (2004) identified the vertical scale as the fundamental element of measuring growth as it places scores from test forms differing in difficulty, but having similar constructs, on a common scale. The process used for associating performance on each test level on a single score scale is called vertical scaling. While the purpose of vertical scaling seems straightforward, the technical aspects of the vertical scaling process involve several considerations. Studies have been conducted on various aspects of vertical scaling.

Several data collection designs are available for linking. Since the common-item non-equivalent groups design does not require that a single group of examinees take multiple test forms, it fits more naturally in the vertical scaling scenarios where multi-level test forms targeting different groups of examinees are used. With this design, the test forms of adjacent levels share a set of common items, and a common scale score can be determined using the common items as an anchor between adjacent test forms. To establish the common scale, the scale linking methods based on the classical test theory (CTT), for example an equipercentile approach, or item response theory (IRT), for example the Rasch model, can be used. Loyd and Hoover (1980) first systematically proposed using the Rasch model with vertical scaling. Several subsequent studies (Patience, 1981; Guskey, 1981; O'Brien & John, 1984) have shown that vertical scaling with the Rasch model performs well in situations where the data fit the model. The assumptions of the Rasch model, such as unidimensionality, need to be checked before fitting the model to the test data (Holmes, 1982; Brogan & Yen, 1983; Loyd & Plake, 1987).

Another technical issue in vertical scaling is the selection of the scale calibration method if the IRT model is applied. The common scale score can be obtained through two scale link approaches. The first approach is via calibrating the item and person parameters of test forms with common items concurrently. The second approach is to calibrate the parameters of the test forms separately, then use some form of scale linking method to put them on the same scale. Hanson & Beguin (2002) found that concurrent estimation generally resulted in better performance than separate estimation when the model is correctly specified. Kim & Cohen (2002) showed that for polytomous data, true polytomous IRT model parameter value 'recovery via concurrent calibration was consistently, albeit only slightly, better than recovery from separate calibration and linking for both item and ability parameters' (p. 39).

Other issues of vertical scaling include the length of the common item set (Raju, Edwards, & Osberg, 1983), the choice of base year (Hendrickson, Cao, Chae, & Li, 2006) and computer software used (Way, Twing, & Ansley, 1988; Custer, Omar, & Pomplun, 2006). Thorough planning needs to be performed before proceeding with vertical scaling since it involves many technical and implementational issues.

Vertical scaling of WIDA ACCESS for ELLs®

The vertical scale of WIDA ACCESS for ELLs® was initially developed using data from a large-scale field test, which was conducted in the fall of 2004. In this section we describe the process of vertical scaling and discuss the results.

Structure of WIDA ACCESS for ELLs®

WIDA ACCESS for ELLs® is a large scale K-12 assessment of academic English language proficiency for English language learners (ELLs) in listening, reading, writing, and speaking. Because it would be unreasonable to develop a single test for all ELLs in the WIDA Consortium, there are several layers of organization to the test series, all of which present issues for the scaling of the tests. In this section, we describe those layers with an eye to the complications that they present. In the subsection 'Developing the logit scale for each domain', we describe how we designed the field test to allow for vertical scaling. In the subsection 'Creating the vertical scale', we describe how the vertical scale was created.

Grade-level clusters. Because an English language proficiency test that is developmentally appropriate for younger students would not be appropriate for older students, and vice versa, WIDA ACCESS for ELLs® is organized into five grade-level clusters, with separate developmentally appropriate tests for each of those clusters. The five clusters are K, 1–2, 3–5, 6–8, and 9–12.[1] To place the tests from different clusters on the same scale, it was necessary to have some items from a lower cluster placed on tests of the next higher cluster, and vice versa. Thus, for example, as we describe in greater detail in the section 'Design of the field test', to enable vertical scaling, items from cluster 1–2 were placed on the 3–5 test forms, and items from cluster 3–5 were placed on the 1–2 test forms.

WIDA's five English language proficiency standards. The goal of WIDA ACCESS for ELLs® is to operationalize WIDA's English Language Proficiency Standards (Gottlieb, 2004; Gottlieb, Cranley, & Oliver, 2007; see also Bailey & Huang, this volume) in an English language proficiency assessment. The WIDA ELP standards emphasize academic English language proficiency and emphasize the development of proficiency in the academic language of five areas:

- Social and Instructional Language (SIL)
- Language of Language Arts (LoLA)
- Language of Math (LoMA)
- Language of Science (LoSC)
- Language of Social Studies (LoSS).

For the Listening test, folders of SIL, LoLA, and LoSC items were used to link across grade-level clusters; while for the Reading test, folders of LoLA and LoMA items were used. In all cases, care was taken to ensure the accessibility of the content of the items in folders that were placed on forms for lower grade-level clusters.

Language domains. Within each grade-level cluster, separate forms were developed for the four language domains: listening, reading, writing, and speaking. By design, the Listening and Reading tests are selected response, while Writing and Speaking are extended constructed response. Thus, the common item nonequivalent groups design was only used for the Listening and Reading forms, and no Writing or Speaking tasks were shared between grade-level clusters. In the case of Writing, this was done to prevent fatigue in students who were already being asked to take a 60-minute writing test. In addition, the Writing tasks needed to be sufficiently complex to allow students in the highest grade level of a cluster (e.g. Grade 8) to demonstrate what they can do with the language. Therefore, Writing tasks from an adjacent grade-level cluster would have been either too easy if from a lower grade-level cluster (e.g. a Grade 1–2 Writing task appearing on a Grade 3–5 cluster form), or too challenging if from a higher grade-level cluster (e.g. a Grade 6–8 Writing task appearing on a Grade 3–5 cluster form). Similarly for Speaking, which is individually administered, adding extended constructed response tasks from any adjacent grade-level cluster would have presented an overly long assessment and one that would have been too easy or too challenging for students. Therefore, it was necessary to find another way to place scores from the test forms in these two domains on a single vertical scale, as described in the subsection ‘Developing the logit scale for each domain’.

Proficiency levels and tiers. The WIDA English Language Proficiency Standards (Gottlieb, 2004; Gottlieb et al., 2007; see also Bailey & Huang, this volume) define six levels of English language proficiency. The first five levels represent increasing English language proficiency development for ELLs; level 6 represents the point at which an ELL’s English language proficiency is such that he or she no longer requires language support to function in a mainstream classroom. Listening and Reading items on WIDA ACCESS for ELLs® are targeted at specific proficiency levels, while each Writing task targets a small range of proficiency levels. Listening and Reading items are grouped in thematic folders of three items at progressively higher proficiency levels. Thus, for example, a Reading folder may have items targeted at levels 1, 2, and 3; 2, 3, and 4; or 3, 4, and 5.

Because the proficiency levels of the WIDA Standards represent a large range of English language proficiency, from initial steps in acquiring English all the way to near-native proficiency, folders containing items at the lower proficiency levels are likely to be insufficiently challenging to higher-proficiency students, while folders at the higher proficiency levels are likely to be frustratingly difficult for lower-level students. In addition, for reliable assessment within each proficiency level, a single form containing items spanning the entire range would be unduly long. Thus, for Writing, Listening, and Reading, there are three overlapping test forms (tiers) within each grade-level cluster. Tier A test forms are designed to allow students at the beginning stages of English language development (i.e. proficiency levels 1 and 2) to demonstrate what they can do, and include Listening and Reading folders with items targeting levels 1, 2, and 3 and

Writing tasks targeted at those levels. Test forms at Tier B are designed for students in the middle range of proficiency (i.e. students at high 2, 3, and low 4) and include folders with items at levels 2, 3, and 4 and Writing tasks targeted at those levels. Finally, test forms at Tier C are designed for students at the higher ranges of proficiency and those eligible for exiting from ELL services, (i.e. high 4, 5, and 6), and include folders with items at levels 3, 4, and 5 and Writing tasks targeted at those levels.

In the field test of WIDA ACCESS for ELLs®, some Tier B folders were included on the Tier A test form, and some Tier C folders were included on the Tier B test. This design allowed for *vertical* linking across the three tiers within each grade-level cluster in addition to vertical scaling across the grade-level clusters. Because of the overlapping nature of the tiers within a cluster, this scaling is fairly straightforward.

Unlike the other domains, there is only one individually administered adaptive Speaking test form per grade-level cluster. Therefore, there is no need for vertical scaling of different speaking tests within a grade-level cluster. The Speaking test is adaptive in the sense that the administrator stops testing in any part of the test if a student demonstrates that the current level of questions is too challenging for him or her (i.e. when a ceiling has been reached).

Design of the field test

As described in the previous section, a vertical scale had to be created for WIDA ACCESS for ELLs® for tests within a grade-level cluster, along with tests across grade-level clusters. In this section we discuss how the field test was designed to allow us to create the scale.

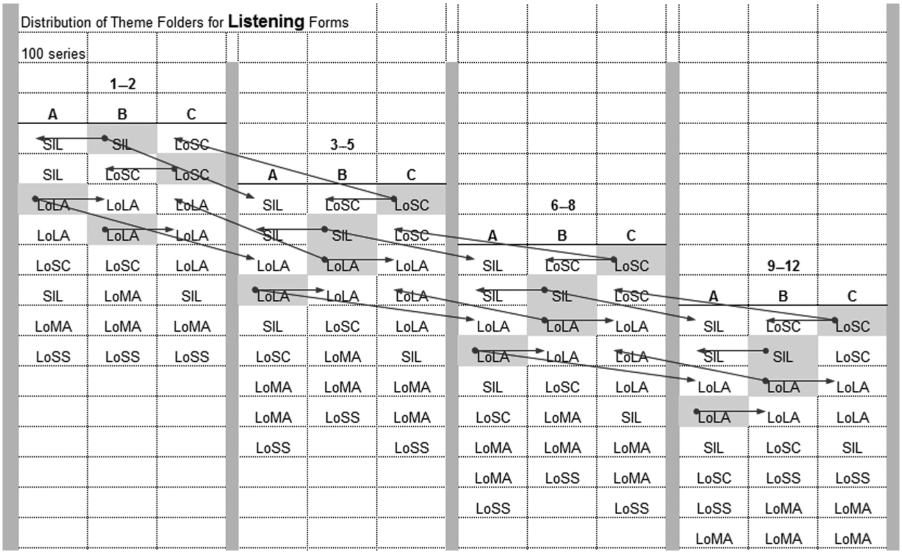
Field test forms. Two forms were created for the Field test, one of which formed the basis of the operational WIDA ACCESS for ELLs®, while the other evolved into a screener test. In this paper, we discuss only the form that later became the operational test.

Listening: As described above, within each grade-level cluster, the Listening test consisted of forms at three tiers: A, B, and C. Each tier contained at least one folder that assessed each of the five standards targeted for that tier and grade-level cluster. In general, following the overall design of the operational test, Tier A Listening tests contained more SIL and LoLA folders, while Tier B and C tests generally contained only one SIL folder targeted to the grade level. Because of the need to create the vertical scale, each tier contained eight or nine folders of three to five Listening items each (as compared to the six folders in the operational form). The distribution of the folders across the grade levels and tiers in the field test form are given in Table 1. Each abbreviation (e.g. SIL) stands for one folder of items targeting that standard. Folders are listed in order of the standards, and not in the order they appeared in the test booklet.

Figure 1 shows how folders were used to link the individual test forms across tiers and grade-level clusters. The shaded cell shows the ‘home’ of that folder; that is, the specifications for which items in that folder were written. The arrows show on which forms that folder also appeared, whether across tiers within a grade-level cluster, or across a grade-level cluster. Thus, for example, a 1–2 SIL Tier B folder was also placed on the 1–2 Tier A test form, and also the 3–5 Tier A test. Similarly, a 3–5 LoSC Tier C folder was also placed on both the 3–5 Tier B test and the 1–2 Tier C test.

Table 1. Listening field test: Thematic folders appearing by standard on each Listening test

Grade-level cluster 1–2			Grade-level cluster 3–5			Grade-level cluster 6–8			Grade-level cluster 9–12		
TierA	TierB	TierC	TierA	TierB	TierC	TierA	TierB	TierC	TierA	TierB	TierC
SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL
SIL	LoLA	LoLA	SIL	LoLA	LoLA	SIL	LoLA	LoLA	SIL	LoLA	LoLA
SIL	LoLA	LoLA	SIL	LoLA	LoLA	SIL	LoLA	LoLA	SIL	LoLA	LoLA
LoLA	LoMA	LoLA	LoLA	LoMA	LoLA	LoLA	LoMA	LoLA	LoLA	LoMA	LoMA
LoLA	LoMA	LoMA	LoLA	LoMA	LoMA	LoLA	LoMA	LoMA	LoLA	LoMA	LoMA
LoMA	LoSC	LoSC	LoMA	LoSC	LoMA	LoMA	LoSC	LoMA	LoMA	LoSC	LoSC
LoSC	LoSC	LoSC	LoMA	LoSC	LoSC	LoMA	LoSC	LoSC	LoSC	LoSC	LoSC
LoSS	LoSS	LoSS	LoSC	LoSS	LoSC	LoSC	LoSS	LoSC	LoSS	LoSS	LoSS
			LoSS		LoSS	LoSS		LoSS			



Distribution of Theme Folders for **Listening** Forms

100 series

1–2		
A	B	C
SIL	SIL	LoSC
SIL	LoSC	LoSC
LoLA	LoLA	LoLA
LoLA	LoLA	LoLA
LoSC	LoSC	LoLA
SIL	LoMA	SIL
LoMA	LoMA	LoMA
LoSS	LoSS	LoSS

3–5		
A	B	C
SIL	LoSC	LoSC
SIL	SIL	LoSC
LoLA	LoLA	LoLA
LoLA	LoLA	LoLA
SIL	LoSC	LoLA
LoSC	LoMA	SIL
LoMA	LoMA	LoMA
LoMA	LoSS	LoMA
LoSS		LoSS

6–8		
A	B	C
SIL	LoSC	LoSC
SIL	SIL	LoSC
LoLA	LoLA	LoLA
LoLA	LoLA	LoLA
SIL	LoSC	LoLA
LoSC	LoMA	SIL
LoMA	LoMA	LoMA
LoMA	LoSS	LoMA
LoSS		LoSS

9–12		
A	B	C
SIL	LoSC	LoSC
SIL	SIL	LoSC
LoLA	LoLA	LoLA
LoLA	LoLA	LoLA
SIL	LoSC	SIL
LoSC	LoSS	LoSS
LoSS	LoMA	LoMA
LoMA	LoMA	LoMA

Figure 1. Listening: Distribution of folders across tiers and grade-level clusters

Reading: The design for the Reading field test paralleled that of the Listening field test, with the exception that LoLA and LoMA folders were consistently used to link tests both across tiers within the same cluster, and across clusters. The distribution of the folders across the grade levels and tiers in the field test form are given in Table 2.

Figure 2 is analogous to Figure 1 and shows how LoLA and LoMA folders were used to link the individual test forms across tiers and grade-level clusters.

Writing: The Writing test within each grade-level cluster-specific form consisted of three tiers: A, B, and C. Each tier's Writing test contained four Writing tasks. Aligned with the ultimate final design of the Writing assessment, each Writing test contained three shorter tasks, allowing students to demonstrate writing proficiency at the levels covered

Table 2. Reading field test: Thematic folders appearing by standard on each Reading test

Grade-level cluster 1–2			Grade-level cluster 3–5			Grade-level cluster 6–8			Grade-level cluster 9–12		
Tier A	Tier B	Tier C	Tier A	Tier B	Tier C	Tier A	Tier B	Tier C	Tier A	Tier B	Tier C
SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL	SIL
SIL	LoLA	LoLA	SIL	LoLA	LoLA	LoLA	LoLA	LoLA	LoLA	LoLA	LoLA
LoLA	LoLA	LoLA	LoLA	LoLA	LoLA	LoLA	LoLA	LoLA	LoLA	LoLA	LoLA
LoLA	LoMA	LoMA	LoLA	LoMA	LoMA	LoLA	LoMA	LoMA	LoLA	LoMA	LoMA
LoMA	LoMA	LoMA	LoLA	LoMA	LoMA	LoMA	LoMA	LoMA	LoMA	LoMA	LoMA
LoMA	LoSC	LoMA	LoMA	LoSC	LoMA	LoMA	LoSC	LoMA	LoMA	LoSC	LoSC
LoSC	LoSS	LoSC	LoMA	LoSS	LoSC	LoSC	LoSS	LoSC	LoSC	LoSS	LoSS
LoSS	LoSS	LoSC	LoSS	LoSS	LoSS	LoSS					
				LoSS							

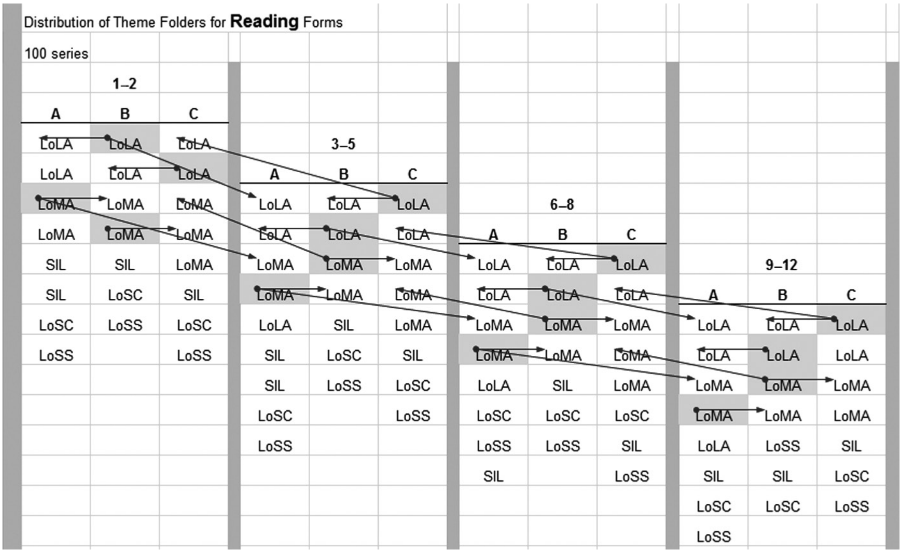


Figure 2. Reading: Distribution of folders across tiers and grade-level clusters

by that tier for the standards SIL, LoMA, and LoSC. It was expected that each of the three shorter tasks would take about 10 minutes to complete. To obtain a longer piece of writing, needed in particular to demonstrate higher proficiency levels, an Integrated Task (IT) was included, for which students were expected to need 30 minutes to complete. Within each grade-level cluster, the IT was designed to elicit writing performances up to proficiency level 5 vis-à-vis LoLA and one or two additional standards (LoSS and SIL). This task was designed to be weighted three times more in the scoring than the shorter Writing tasks. All Writing tasks are scored using a common rubric that reflect the performance level descriptors of the WIDA Standards across all grade-level clusters.

In the field test, within each grade-level cluster, there was a horizontal overlap across the three tiers, as a common IT appeared on the form for all three tiers. In addition, Tiers B and C shared a common SIL Writing task.

The distribution of the Writing tasks within the grade-levels clusters and across tiers in the field test form are given in Table 3.

Speaking: The Speaking assessment was designed for individual, one-on-one administration. To reduce administration time as much as possible, an adaptive format was chosen. Table 4 shows the basic design used in the field test and intended for use in the operational test. One form of the assessment was created for each grade cluster, and all the grade clusters used this same design.

The Speaking assessment consists of three thematic folders: SIL, LoLA/LoSS and LoMA/LoSC. Students are first administered the Level 1 speaking task from the SIL thematic folder. The Test Administrator (TA) continues administering the tasks within that folder until exhausting all the tasks within a folder or reaching a task for which, in the TA's judgment, the student is unable to meet the task expectations. At that point, the TA moves to the Level 1 task from the next thematic folder.

Following the above guidelines for administration, the Speaking assessment becomes adaptive to the current proficiency level of each student. The student currently at the lowest level may be administered a minimum of three tasks (i.e. the Speaking task at proficiency Level 1 from each of the three thematic folders), while the student at the highest proficiency level may be administered all 13 Speaking tasks. For students who complete all 13 tasks, the test administration typically takes about 15 minutes.

Field test sample. For the group-administered sections of the field test (i.e. Listening, Reading, and Writing), a total of 3633 students from eight of the WIDA Consortium

Table 3. Writing field test

Writing field test				
Tier A	SIL	LoMA	LoSC	IT_{ABC}
Tier B	SIL_{BC}	LoMA	LoSC	IT_{ABC}
Tier C	SIL_{BC}	LoMA	LoSC	IT_{ABC}
Time Allowed	10 mins	10 mins	10 mins	30 mins
Weighting	1	1	1	3

Note: Subscripted letters indicate tasks that were used across multiple tiers (e.g., the same SIL task was used for Tiers B and C).

Table 4. Speaking form design

	Speaking form design				
	Proficiency Level 1	Proficiency Level 2	Proficiency Level 3	Proficiency Level 4	Proficiency Level 5
SIL	Speaking task	Speaking task	Speaking task		
LoLA/LoSS	Speaking task	Speaking task	Speaking task	Speaking task	Speaking task
LoMA/LoSC	Speaking task	Speaking task	Speaking task	Speaking task	Speaking task

Table 5. Distribution of field test sample by grade-level cluster

	Grade-level cluster			
	1–2	3–5	6–8	9–12
Count	680	1,008	915	1,030

states participated in the field test. Table 5 shows the number of students per grade-level cluster. Within each grade-level cluster, students were divided roughly equally among the three tiers. This is an adequate sample size for this type of field testing and certainly sufficient for the analyses planned for each of the grade clusters.

The Speaking test was individually administered by trained raters in two of the WIDA Consortium states. A total of 523 students, who had also taken the group administered section, participated in the field test of the speaking section.

Creating the vertical scale

Creating and calibrating the data matrix. In preparation of a concurrent calibration of all the items onto a single scale, the original responses were entered into one large, two-dimensional data file, with items in the columns and test takers in the rows. The cells at the intersection of each row with each column contained that test taker's response on that item. This data file included all the items in each domain, from proficiency level 1 through proficiency level 5, and all the students in the sample, from 1st-grade students in the 1–2 cluster to 12th-grade students in the 9–12 cluster. Of course, not every student took every item, so many of the cells in this matrix of scores were empty. But because of the common item nonequivalent groups design, booklets were linked together by folders of linking items; therefore, every student took items from more than one tier, and students in Tier A and Tier C took items from more than one grade-level cluster. Between any two forms that were linked, at least two out of the seven or eight folders (i.e. 25% to 29% of the items) were common between them (see Figures 1 and 2).

The computer program used to score and analyze this data, Winsteps (Linacre, 2006), is very robust to missing data and is able to analyze such a matrix quite adequately. In addition, because students were taking 'off-target' items (i.e. items from other tiers or grade-level clusters that might have been too hard for them – thus encouraging guessing – or too easy for them – thus encouraging inattentiveness), student responses on items that, according to the analyses, were much too easy or much too hard for the students were marked as missing, using the analysis control functions provided by Winsteps. Specifically, responses to an item by students whose ability level was three logits higher than the item difficulty were excluded (i.e. the item was treated as 'too easy'), as were responses to an item by students whose ability level was two logits lower than the item difficulty (i.e. the item was treated as 'too hard') in the concurrent calibration. Such responses have the potential of negatively influencing the linking process. In this way, it was ensured that only data from students across tiers and grade levels for whom the items were genuinely appropriate was used in the analysis. In addition, very tight convergence criteria were used in calibrating the data. Winsteps has two control functions for establishing convergence criteria: LCONV, which sets the logit change at the convergence, and RCONV, which sets the score residual at convergence. The default value for LCONV is 0.005; we set it at 0.001. Similarly, the default value of RCONV is 0.1; we set it at 0.01. Using such tight convergence criteria is required because of the 'long chain' (i.e. stretching across tests from grade-level clusters 1–2, 3–5, 6–8, and 9–12) formed in the linking design.

Developing the logit scale for each domain. For the Listening and Reading items, the difficulty of all the items was calibrated onto one common scale through a concurrent calibration of the single data matrix (one for Listening and one for Reading). Concurrent calibration accomplishes two important things: first, it creates a single logit scale for all the items (from Grade 1 to Grade 12), which forms the basis for the reporting scale; and second, it links the test forms by putting all the items, from every booklet, on the same scale. Likewise, it also puts the ability measures of the students from Grades 1 to 12 on the same scale. A detailed description of the scaling procedure can be found in Kenyon (2006).

In the case of Writing and Speaking, a slightly different procedure was used to put examinee performances onto a common logit scale across the grade-level clusters. Although a single rating scale was used across the grade-level clusters for Writing, there were no common tasks linking the writing booklets across the grade clusters. Because of the high degree of relationship between reading and writing skills, performances on the Reading items, which were already calibrated onto a single scale across the grade-level clusters, were used as a scaling test to make the adjustments to the separate calibrations within the four grade-level clusters for Writing. To do this, a single data file was created with student responses to both the Writing tasks and the Reading items. As the difficulty level of the Reading items had already been placed on a logit scale extending across the grade-level clusters, in this calibration the difficulty values of the Reading items were anchored (i.e. fixed to the values derived from their concurrent calibration, just described). The Writing tasks (and the scale steps) were unanchored and thus were calibrated in this run onto that same scale. Table 6 shows a summary of the Rasch fit statistics for the writing tasks calibrated to the anchored reading tasks. The results show that the vast majority of writing items fit the Rasch measurement model combining reading and writing tasks,

Table 6. Writing: Distribution of Mean-Square Fit Statistics

Range of Mean-Square Fit Statistic	INFIT	OUTFIT
> 2.0	N = 1	N = 1
"distorting or degrading measurement"	% = 2.0%	% = 2.0%
> 1.5–2.0	N = 6	N = 7
"unproductive but not degrading"	% = 11.8%	% = 13.7%
0.5–1.5	N = 44	N = 43
"productive for measurement"	% = 86.3%	% = 84.3%
< 0.5	N = 0	N = 0
"less productive but not degrading"	% = 0%	% = 0%
Total	N = 51	N = 51
	% = 100%	% = 100%

implying that the items from the two domains were measuring some common construct. Only one writing task (2% of the total) was in the range that Linacre (2002) labels ‘distorting or degrading measurement’.

This joint calibration of Reading items and Writing tasks had the effect of putting all the difficulty values of the Writing items on the same scale as the Reading items; that is, putting all the Writing tasks onto one common logit scale. The practical outcome of this process was that task-difficulty logit values for Writing tasks on each grade-level cluster become increasingly higher from grade-level cluster to grade-level cluster. In other words, the Grades 1–2 task-difficulty values for the Writing tasks became lower than those for the Grades 3–5 Writing tasks, which in turn were lower than those for the Grades 6–8 tasks, and so on.

To calibrate the difficulty of the Speaking tasks on a common vertical scale across the grade-level clusters, the same process was used, with the only difference being that the precalibrated performances on the listening items were used as the scaling test to link the Speaking items across grade-level clusters. A detailed description of the scaling procedure for speaking can be found in Kenyon (2006). Table 7, analogous to Table 6, shows a summary of the Rasch fit statistics for the speaking tasks calibrated to the anchored listening tasks. Again, the results show that using Linacre’s guidelines, no items were unproductive, degrading, or distorting to the measure of a common underlying construct when the items from the two domains were combined into one measure.

From logits to scale scores and proficiency scores. After developing the logit scale across the grade-level clusters, a reporting scale was created for each of the four domains (Listening, Reading, Writing, and Speaking). In addition, four composite scale scores were developed, as detailed below. Converting a logit scale to a reporting scale involves two steps: multiplying the logits by a spacing factor, and then adding an additive factor. Logit scales typically have small units that require decimals for reporting and center on 0; a typical logit scale might range from −4 to 4. The purpose of the spacing factor is to get rid of the decimal. Thus, for a logit scale running from −3.99 to 3.99, we can multiply the logits by 100 to create a scale from −399 to 399. Then, to ensure that all scale scores are greater than 0, we can add 400, an example additive factor, giving us a scale that runs from 1 to

Table 7. Speaking: Distribution of mean-square fit statistics

Range of mean-square fit statistic	INFIT	OUTFIT
> 2.0	N = 0	N = 0
‘distorting or degrading measurement’	% = 0%	% = 0%
> 1.5–2.0	N = 0	N = 0
‘unproductive but not degrading’	% = 0%	% = 0%
0.5–1.5	N = 51	N = 39
‘productive for measurement’	% = 98%	% = 75%
< 0.5	N = 1	N = 13
‘less productive but not degrading’	% = 2%	% = 25%
Total	N = 52	N = 52
	% = 100%	% = 100%

799 and is centered on 400. In the context of the WIDA ACCESS for ELLS® test, it was desired that these two factors be calculated with an eye to ensuring that the scale could be interpreted by all stakeholders.

For WIDA ACCESS for ELLs®, the reporting scales were designed to range from 100 to 600. Since across the four domains the logit scales were independent of each other, it was decided to have a common frame of reference across the four scales by setting the center point of each scale (i.e. 350) to be the cut score, set by the Consortium's standard setting process (Kenyon, 2006), between proficiency Levels 3 and 4 for the 3–5 grade-level cluster.

To develop the scale across the form domains, the logit scale for the reading test was chosen, as it had the largest range of scores. Because the four separate scale scores would be used to form composites, it was also necessary to ensure that relative weighting of each domain scores would be equal and not a function of their variation. To avoid this, we followed a procedure that in effect created a standardized z-score for each domain. Thus, the standard deviation of each logit scale was calculated, and the ratio of the standard deviation for the logit scale for each domain to the standard deviation of reading was calculated (i.e. for reading it was 1.00). Next, the spacing factor for reading was determined, ensuring the widest possible spread of points within the 100–600 reporting scale range with the score of 350 at the logit value of the cut between Levels 3 and 4 for grade-level cluster 3–5. The spacing factor for reading turned out to be 26. To determine the spacing factor for the other domains, 26 was multiplied by the ratio of the standard deviation of that domain to the standard deviation of reading. Finally, the additive factor was calculated for each scale such that the chosen cut would equal 350.

The four composite scores that are calculated for WIDA ACCESS for ELLs® are Oral Language, Comprehension, Literacy, and Overall. The relative weights of each domain in those scores were set by policy of the WIDA Consortium and are as follows:

- Oral Language Composite: 0.5 * Listening + 0.5 * Speaking
- Comprehension Composite: 0.3 * Listening + 0.7 * Reading
- Literacy Composite: 0.5 * Reading + 0.5 * Writing
- Overall Composite: 0.15 * Listening + 0.35 * Reading + 0.35 * Writing + 0.15 * Speaking

As mentioned earlier, scale scores for the four domains, along with the composite scores, are interpreted in terms of the WIDA ACCESS for ELLs® proficiency levels based on cut scores set in each domain by the WIDA Consortium's standard setting process. Cut scores for the composite scores were determined by solving the equations above using the corresponding cut score for each of the domains. For example, the cut score for proficiency Levels 1 and 2 for grade 2 are 247 for Listening, 267 for Reading, 251 for Writing, and 286 for Speaking. Using the formula above, we calculate the Overall Composite cut score as 0.15 * 247 + 0.35 * 267 + 0.35 * 251 + 0.15 * 286, which works out to a scale score of 261.

In addition to the proficiency level, proficiency level scores are determined for each of the domains and the composite scores. A proficiency level score consists of a two-digit decimal number (e.g. 4.5). The first digit represents the student's overall language proficiency level range based on the student's scale score. A score of 4.5 indicates that the student is in language proficiency Level 4. The number to the right of the decimal is an indication of the proportion of the range between cut scores that the student's scale score represents. A score of 4.5 indicates the student's scale score is halfway between the cut scores for Levels 4 and 5. Thus, for example, the 3/4 cut score for Writing for grade 2 is 320, while the 4/5 cut score is 348. A scale score of 334 is halfway between those two cut scores; therefore, a second grade student who receives a scale score of 334 on the Writing test would receive a proficiency level score of 4.5 for Writing.[2]

Evaluation of the effectiveness of the vertical scaling of WIDA ACCESS for ELLs®

While the adequacy of the vertical scaling of WIDA ACCESS for ELLs® was examined to a certain extent using the field test data (see Kenyon, 2006), the real test of its effectiveness comes when longitudinal data on individual students are used to study growth over time. The WIDA Consortium has used the WIDA ACCESS for ELLs® vertical scale in several growth trend analyses, capitalizing on the uniqueness resident in vertical scaling. As stated earlier, one of the requirements in NCLB is to establish AMAOs for district accountability purposes. Three types of AMAOs are mandated: growth (AMAO 1), attainment (AMAO 2), and adequate yearly progress (AMAO 3). Since vertical scales are ideal for tracking growth, they can be used to support the development of AMAO 1 targets. Cook, Boals, Wilmes, and Santos (2008) analyzed data from three WIDA Consortium states to provide recommendations regarding AMAO 1 targets. Figure 3 (adapted from Cook et al., p. 15, fig. 6) displays growth characteristics of WIDA ACCESS for ELLs® used for their AMAO 1 analyses.

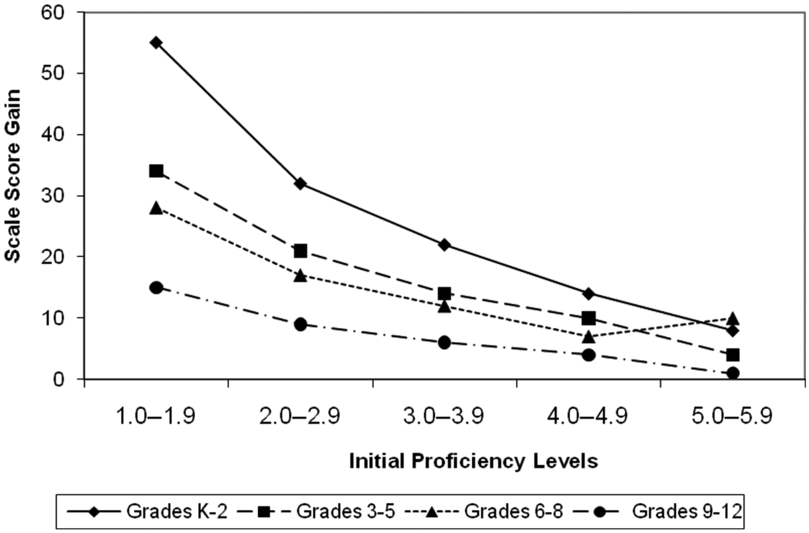


Figure 3. Average scale score gain by grade band and proficiency level

Table 8. Average scale score gain by grade cluster and proficiency level band

Grade Cluster	Band				
	1.0–1.9	2.0–2.9	3.0–3.9	4.0–4.9	5.0–5.9
K–2	55	32	22	14	8
3–5	34	21	14	10	4
6–8	28	17	12	7	10
9–12	15	9	6	4	1

Since most states, districts, and schools use the overall composite score to make decisions about students' levels of English language proficiency and their need for support services, this analysis focuses on that score. The y-axis of Figure 3 displays the annual growth in overall composite scale score units and the x-axis displays WIDA proficiency level bands. Table 8 shows the average gain by proficiency band and grade cluster. Notice that students initially scoring at the lower proficiency levels have a higher composite proficiency growth rates, in scale score units, than students initially scoring at higher proficiency levels. Similarly, students at lower grades grow faster than students in higher grades. This observed characteristic fostered the principle outlined in Cook et al.'s (2008) paper 'lower is faster, higher is slower.' That is, lower proficiency levels or lower grades (i.e. younger students) will demonstrate growth at a faster rate than students in higher proficiency levels or higher grades. This non-linear growth rate is consistent with what has been reported in child second language acquisition literature (e.g. Hyltenstam & Abrahamsson, 2003; Collier, 1995; Long, 2003). This type of growth rate is also well documented in foreign language learning and has led to the famous 'inverted pyramid' description of foreign language growth used by the American Council on the Teaching of Foreign Languages (Swender, 1999). It is also to be expected in an assessment of academic English language proficiency; at higher grades the breadth and depth of academic content is much greater than in lower grades. The outcome of these analyses, based on WIDA ACCESS for ELLs® vertically scaled scores, provides some evidence in support of their use. It could be argued that the similarity of the observed non-linear characteristics seen from this analysis to other similar analyses in the second language acquisition research support the use of the vertical scaling on WIDA ACCESS for ELLs®.

The WIDA Consortium continues to support states' use of growth information based on these vertically scaled scores and has recently published a bulletin providing the most recent picture of growth across WIDA states (Cook, 2009a). Strategies for using this information to support students' acquisition of academic English are also discussed in this bulletin.

Another way the vertically scaled ACCESS scores have been utilized by states is in an investigation to answer the question: at what point should English language learners be exited from ELL services? This research examines the relationship between student performances on state reading and mathematics content assessments and on WIDA ACCESS for ELLs®. Using a decision theoretic approach commonly employed in bias reviews or standards setting, Cook (2009b) has shown that across grades there is a relationship between the student English language proficiency level as measured by WIDA ACCESS for ELLs® and the likelihood of scoring at the 'proficient' level in state content tests. However, this relationship only holds up to overall English language proficiency levels between 4.5 and 5.0, depending on state, grade, and subject. Above this range, English language proficiency no longer appears to play a decisive role in a student's classification as 'proficient' on the state content test, implying that for such a student the state content assessment may be measuring content knowledge with less interference from the student's current level of English language proficiency.

Conclusion

In this paper, we have reviewed some technical aspects of vertical scaling based on a brief literature review. We have also presented one example of vertical scaling used in a large-scale English language proficiency test spanning grades K to 12, along with analyses that provide some confirmatory evidence in support of the success of the scaling using longitudinal data. As Young (2006) writes, 'Vertical scaling is an intricate measurement process. Growth definitions, scale assumptions, data collection designs, and linking methods combine to produce vertical scale [*sic*] in ways that are not always clear. However, what is clear is the need to carefully document the construction of any vertical scale' (p. 484). Kenyon (2006) gives a more complete documentation of the issues and complexities in creating a vertical scale spanning grades K to 12 for an English language proficiency test in four language domains. In this paper we have attempted to illustrate some of the issues that are especially relevant to a broader audience of language testers.

Notes

1. The format of the Kindergarten version of the test differs from that in the other grade clusters in that all domains are administered one-on-one. However, it shares items and tasks in common with the grade 1–2 test forms. Thus, performances on the Kindergarten form were placed on the K-12 scale using a common-item approach linking them to performances on the grades 1–2 test forms, domain by domain. This relatively straightforward approach is not discussed further in this paper.
2. Note that, unlike the scale scores, the proficiency level scores are not linear. That is because the distance between cut scores is not constant, even within a domain and grade-level cluster. For example the distance between the 2/3 and 3/4 cuts for grade 2 Writing is 35 scale score points, while the distance between the 3/4 and 4/5 cuts for the same grade and domain is 28 scale score points. Therefore an increase of 14 scale score points for a student in proficiency level 3 would correspond to an increase of 0.4 proficiency level points, while a similar increase of 14 scale score points for a student in proficiency Level 4 would correspond to an increase of 0.5 proficiency level points.

References

Brogan, E. D., & Yen, W. M. (1983). Detecting multi-dimensionality and examining its effects on vertical equating with the Three-Parameter Logistic Model. Paper presented at the Annual Meeting of the American Educational Research Association, Montreal.

Collier, V. (1995). Acquiring a second language for school. *Directions in Language & Education, 1*(4). Retrieved May 19, 2008, from www.ncela.gwu.edu/pubs/directions/04.htm.

Cook, H. G. (2009a). *WIDA Focus on Growth*. WIDA Focus on Series, vol 1: 1. Retrieved May 19, 2008, from www.wida.us/ACCESSTraining/Focus/on_Growth.pdf.

Cook, H. G. (2009b). *Annual Measurable Achievement Objective # 2 (AMAO 2): A Report on a Meeting Comparing the Kentucky Core Content Test (KCCT) and ACCESS for ELLs™ to Establish an AMAO 2 Exit Criterion.* Kentucky Department of Education.

Cook, H. G., Boals, T., Wilmes, C., & Santos, M. (2008). *Issues in the development of annual measurable achievement objectives for WIDA consortium states* (WCER Working Paper No. 2008-2). Madison, WI: University of Wisconsin–Madison, Wisconsin Center for Education Research. Retrieved May 19, 2008 from www.wcer.wisc.edu/publications/workingPapers/papers.php.

Custer, M., Omar, M. H., & Pomplun, M. (2006). Vertical scaling with the Rasch model utilizing default and tight convergence settings with WINSTEPS and BILOG-MG. *Applied Measurement in Education*, *19*, 133–149.

Gottlieb, M. (2004). *English language proficiency standards for English language learners in kindergarten through Grade 12: Framework for large-scale state and classroom assessment*. Madison, WI: WIDA Consortium.

Gottlieb, M., Cranley, M. E., & Oliver, A. (2007). *Understanding the WIDA English language proficiency standards: A resource guide*. Madison, WI: WIDA Consortium.

Guskey, T. R. (1981). Comparison of a Rasch model scale and the grade equivalent scale for vertical equating of test scores. *Applied Psychological Measurement, 5*(2), 187–201.

Hanson, B. A. & Beguin, A. A. (2002). Obtaining a common scale for IRT item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3–24.

Hendrickson, A. B., Cao, Y., Chae, S., & Li, D. (2006). *Effect of Base Year on IRT Vertical Scaling from the Common-Item Design.* Paper presented at the Annual Meeting of the National Council for Measurement in Education, San Francisco, CA.

Holmes, S. E. (1982). Unidimensionality and vertical equating with the Rasch model. *Journal of Educational Measurement*, *19*(2), 139–147.

Hyltenstam, K., & Abrahamsson, N. (2003). Maturational constraints in SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 539–588). Malden, MA: Blackwell.

Kenyon, D. M. (2006). *Development and Field Test of ACCESS for ELLs®* (WIDA Consortium Technical Report No. 1). Retrieved June 23, 2009, from www.wida.us/assessment/ACCESS/TechReports/Technical%20Report%201.pdf.

Kim, S.-H., & Cohen, A. S. (2002). A comparison of linking and concurrent calibration under the graded response model. *Applied Psychological Measurement, 26*(1), 25–41.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating: Methods and Practices* (2nd ed.). New York: Springer-Verlag.

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, *16*(2), www.rasch.org/rmt/rmt162f.htm.

Linacre, J. M. (2006). *Winsteps Rasch measurement.* [Computer Software]. Chicago, IL: Winsteps.

Long, M. H. (2003). Stabilization and fossilization in interlanguage. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 487–535). Malden, MA: Blackwell.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch Model. *Journal of Educational Measurement, 17*, 179–193.

Loyd, B. H., & Plake, B. S. (1987). *Vertical equating: Effects of model, method, and content domain.* Paper presented at the Annual Meeting of American Educational Research Association, Washington, DC.

No Child Left Behind Act of 2001. Public Law No. 107–110, § 115 Stat. 1425 (2002).

O'Brien, M. L., & John, D. (1984). Applying and evaluating Rasch vertical equating procedures for out-of-level testing. Paper presented at the Annual Meeting of the Eastern Educational Research Association, West Palm Beach.

Patience, W. M. (1981). A Comparison of latent trait and equipercentile methods of vertically equating tests. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Los Angeles.

Raju, N. S., Edwards, J. E., & Osberg, D. W. (1983). The effect of anchor test size in vertical equating with the Rasch and three-parameter models. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Swender, E. (1999). *ACTFL OPI Tester Training Manual – Revised 1999.* New York: ACTFL Publication.

Way, W. D., Twing, J. S., & Ansley, T. N. (1988). A comparison of vertical scalings with the three-parameter model using logist and bilog and two different calibration procedures. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans.

Young, M. J. (2006). Vertical scales. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 469–485). Mahwah, NJ: Lawrence Erlbaum.