WIDA™ RESEARCH REPORT

EXAMINING HOW TO ESTABLISH ENGLISH PROFICIENCY USING ALTERNATE COMPOSITE SCORES IN A WIDA STATE:

A Logistic Regression Approach

PREPARED BY
H. Gary Cook, Ph.D., Wisconsin Center for Education Research,
University of Wisconsin-Madison

APRIL 2013

WCER
WISCONSIN CENTER FOR EDUCATION RESEARCH

Thanks to Robert Linquanti for his excellent comments.
However, any errors or misrepresentations in this document
are solely the author's.

World Class Instructional Design and Assessment (WIDA) advances academic language development and academic achievement for linguistically diverse students. WIDA was formed as the result of a federal grant to comply with the requirements of the No Child Left Behind Act. It is a consortium of states and districts working together to promote achievement of English language learners. The organization has created a comprehensive system that includes English Language Development Standards, Spanish Language Development Standards, English language proficiency assessments, professional development for educators of ELLs, and research on all aspects of English language learning.

RESEARCH

WIDA's Reseasrch Department seeks to provide timely, meaningful, and actionable research that promotes educational equity and academic achievement for linguistically and culturally diverse students. Its annual research agenda is developed under the guidance of the WIDA Consortium Board Research Subcommittee and includes topics in the areas of academic language, standards, professional learning, and policy.

Goal of Analyses

The goal of analyses presented here is to identify a procedure for creating alternate composite scores on English language proficiency assessments without using all four domain test scores (i.e., listening, speaking, reading and writing). Acceptable alternate composite scores could then be used for accountability purposes for English learners (ELs) who have a documented disability that prevents them from participating on all four domain tests (e.g., the listening portion of the ELP test for ELs who are hearing impaired). The following is a list of assumptions made to create alternate composite scores.

Assumptions

1. Proficiency on state reading and mathematics content assessments are reasonable outcome measures for determining English-language proficiency alternate composite scores. The approach is consistent with the definition of an EL in federal statute §9101(25).

2. Equally weighted composite scores are reasonable measures to represent English-language proficiency[1].

3. ELs with disabilities do not perform differently than their non-disabled EL counterparts when comparing English-language proficiency and academic content performance levels[2].

4. The estimated probability of 0.50 using logistic regression models is a reasonable point for establishing English-language proficiency for an alternate composite score[3].

5. Comparing rescaled R^2 (or equivalent statistics) is an acceptable method for comparing alternate composite score models' goodness of fit.

Data Used

Data used for analyses presented here come from a WIDA state with approximately 50,000 English learners from Kindergarten to twelfth grade. Only third grade data from the 2009-2010 school year are used for analyses. In total, 4,921 EL students had ACCESS for ELLs® overall composite scores. Of that original number, 4,881 (99%) ELs had both ACCESS and state mathematics assessment scores, and 4,660 (95%) had both ACCESS and state reading comprehension scores.

[1] WIDA's ACCESS for ELLs® assessment does not equally weight all domains. Literacy is weighted by 70%. For illustrative purposes and ease in interpretation an equal weighting methodology is used here.

[2] Also implicit in this assumption is that an EL's disability does not interfere with their opportunity to engage in the academic content assessment item/task and demonstrate what they know with appropriate accommodations.

[3] This assumption is amended somewhat for alternate composite models with less than optimal goodness of fit.

Analyses

Based on the assumptions made in the previous section, fourteen logistic models were examined for ELs at third grade in a WIDA state, designated State A. All models have the following form,

$$\pi_k = \frac{1}{1 + e^{-(\alpha+\beta x)}},$$

where π_k is the probability of an EL being proficient on State A's content test *k*, given alternate composite score *x*, and α and β are logistic parameter estimates for the intercept and scope respectively. Most available statistical software provides goodness of fit statistics for logistic regression models. The analyses reported here were conducted using SAS[4], and the goodness of fit statistic used is the rescaled R^2 estimate,[5] which is analogous to R^2 in linear regression models.

The fourteen alternate composite models examined [i.e., (reference model + 6 alternatives) x (reading and mathematics) =14] each have different domain score combinations, such that,

Reference Model: Listening, Speaking, Writing and Reading = LSWR,

Model 1: Speaking, Writing, Reading = SWR,

Model 2: Listening, Writing, Reading = LWR,

Model 3: Listening, Speaking, Reading = LSR,

Model 4: Listening, Speaking, Writing = LSW,

Model 5: Writing, Reading (Literacy) = WR, and

Model 6: Listening, Speaking (Oral) = LS.

[4] The data analysis for this paper was generated using SAS software, Version 9 of the SAS System for Windows. Copyright © 2012 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

[5] http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect065.htm

Results

Each alternate composite model was calculated by averaging the relevant domain scale score values. Table 1 displays the rescaled R^2 values from logistic regressions between each alternate composite score model and State A's mathematics and reading content assessment proficient scores. The LSWR model is used as a reference (i.e., this "alternate" composite score represents an overall composite combination that equally weights all domain scale scores). Many states (e.g., WIDA Consortium states) differentially weight domains in creating overall composite scores. Since analyses shown here are for illustrative purposes, differential weighting was not applied.

TABLE 1: Rescaled R^2 & Proportional Relationships for a Variety of Logistic Regression Models Examining Different Alternate Composite Scores for Grade 3 in State A

Model	Domains	Rescaled R^2 Mathematics	Rescaled R^2 Reading	Proportional Relationship to an all-domain composite		
	LSWR	25.6%	37.4%	Math	Read	Avg.
1	SWR	22.7%	35.2%	89%	94%	91%
2	LWR	27.5%	40.7%	107%	109%	108%
3	LSR	22.2%	30.5%	87%	82%	84%
4	LSW	23.7%	34.4%	93%	92%	92%
5	WR	25.3%	41.5%	99%	111%	105%
6	LS	18.6%	24.6%	73%	66%	69%

The first four columns in Table 1 display information on the R^2 values for each of the fourteen alternate logistic models. The rescaled R^2 values for the reference model are 25.6% and 37.4% for mathematics and reading, respectively. That is, 25.6% of the variability in ELs' likelihood of being proficient on State A's mathematics test can be attributed to the overall composite score (LSWR). Likewise, 37.4% of the variability in the likelihood of being proficient on State A's reading comprehension test can be attributed to the overall composite score. When an alternate composite score is created, absent of the listening domain scale scores (i.e., Model 1, SWR), the R^2 decreases to 22.7% for mathematics and 35.2% for reading. The proportional relationship between the R^2 for the alternate composite, absent of listening, and the overall composite is 89% (22.7% ÷ 25.6%). The latter three columns in Table 1 show the proportional relationship between alternate composite models and the overall composite. The last column is the average of the mathematics and reading R^2 estimates. Notice that most of these proportions are above 80%. Some are above 100%. For example, the R^2 for the literacy alternate composite model (WR) is a *better predictor* for proficiency in reading comprehension than the overall composite, i.e., 41.5% versus 37.4%. This might be expected since both the literacy composite and reading comprehension test are assessing very similar constructs. The lowest proportional relationship between the overall composite is found in Model 6, the oral composite model (LS).

How might these results be used to

1. determine if an alternate composite score model has sufficient goodness of fit compared to the overall composite model to be used, and

2. create an alternate composite score that might be used for accountability purposes, for AMAOs?

To illustrate how one might respond to both questions, two alternate composite models are compared: the literacy composite (WR) and the oral composite (LS). These models were chosen because one common concern expressed by states is that deaf or blind EL students cannot be assessed on certain domain assessments because of their disability. That is, deaf ELs are unable to meaningfully participate on the oral section of the language proficiency test, and blind ELs are unable to participate on the literacy section the test.

Goodness of fit

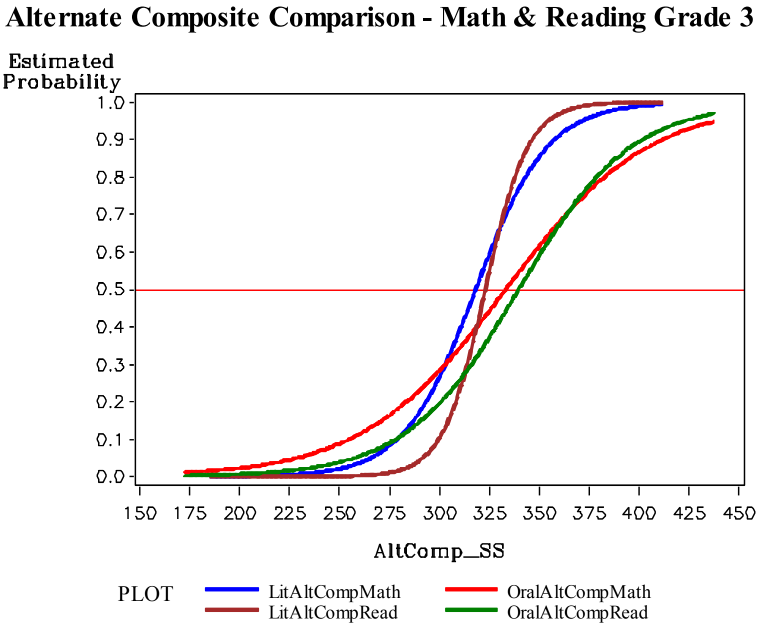
As seen in Table 1, the literacy composite model has a high degree of association with the goodness of fit of the overall composite (LSWR). For mathematics, the literacy composite provides an almost identical R^2 value (WR=25.3% vs. LSWR=25.7%). There is a 99% overlap between models. For reading, the literacy score has a higher R^2 value (WR=41.5% vs. LSWR=37.4%). The literacy composite of the ELP test is a comparable, if not a better, predictor of the variability in the likelihood of being proficient on State A's content assessments attributable to English language proficiency, and could easily be used as an alternate composite score and not affect the predictive relationship between English language proficiency (in this case literacy) and academic content proficiency.

The comparative predictive relationship between the oral alternate composite score and the overall composite is much less than that of literacy, with the mathematics R^2 being 18.6% and reading comprehension R^2 being 24.6%. The oral composite's goodness of fit only overlaps by 66% in mathematics and 73% in reading comprehension compared to the overall composite. This is noticeably less than literacy or any other alternate composite combination. The oral composite is the least predictive of the variability in the likelihood of being proficient on content assessments of all alternate composite models.

Creating a Proficiency Score for Alternate Composite Models

Assumption 4, above, states that an estimated probability of p=0.50 is representative of English language proficiency (i.e., an EL student at that ELP level is equally likely to meet the academic content test's proficient performance standard as not). Logistic regression models provide response probabilities for predicted scores. Figure 1 plots the response probabilities for both the literacy and oral alternate composite scores for the mathematics and reading content assessments.

FIGURE 1: Logistic Regression Curves for the Literacy and Oral Alternate Composite Scores and the Estimated Probability of being Proficient on State A's Mathematics and Reading Comprehension Tests



In Figure 1 the x-axis displays the alternate composite scale score range and the y-axis shows the estimated probabilities for being proficient. The red, horizontal line identifies p=0.50. The legend shows the alternate composite/content curves. Thus "LitAltCompMath" is the predictive probability curve for the literacy alternate composite score and proficiency in mathematics. The other alternate composite/content curves are similarly labeled. *Note that comparing the position of each alternate composite's curve relative to the other is not meaningful.* Each domain test is scaled separately; thus, each alternate composite combination has unique means and variances. They only share the same minimum and maximum scores. What is relevant is the *slope of each curve*. The literacy composite has a much steeper slope. This means that for any one scale score unit change, the probability of being proficient on State A's mathematics or reading comprehension test is greater for the literacy composite than for the oral composite. This observation is also reflected in each composite's rescaled R^2 values. With data used to create these curves, scale score values representing a predicted probability at p=0.50 can be identified.

TABLE 2: Alternate Composite Scale Scores for 0.50 and for 0.60 Probability from Logistic Regression Models for State A for Mathematics and Readings Assessments at Grade 3

Alternate Composite Model	Mathematics	Reading
#5: Literacy (WR) at p=0.50	318	323
#6: Oral (LS) at p=0.50	333	339
#6: Oral (LS) at p=0.60	348	351

For the literacy composite, a scale sore representing a p=0.50 probability is 318 for mathematics and 323 for reading. For the oral composite the scale score values are 333 and 339 for mathematics and reading. *Again, comparing score values between the literacy and oral composites is not meaningful or useful.* For example, the difference in scale scores between the literacy and oral composite does NOT mean that you have to have more "oral language" than "literacy" to be proficient. Note that a p=0.60 probability row is added for the oral composite. As seen in Table 1, the predictive capacity of the oral composite score is much less than the overall composite score. Because of this limited, relative predictability, a score that yields a higher probability is recommended. Stated differently, the oral composite score has a greater degree of "noise" compared to the overall composite–or the literacy composite–in determining whether 3rd grade ELs have an equal likelihood of being proficient on State A's mathematics or reading comprehension tests. Thus, a score yielding a higher threshold probability is recommended.

The following illustrates a sample rule of thumb: If the proportional relationship between an alternate composite and an overall composite score is 80% or higher, the p=0.50 probability should be used. If the proportional relationship between an alternate composite and overall composite is between 60% and 79%, a p=0.60 probability should be used. If the proportional relationship is less than 60%, the alternate composite score should not be used. Note that this sample "rule of thumb" is arbitrary. Why 80%? Why not 90% or 75%? Why p=0.60? Acknowledging that the recommendation made here is ultimately a judgmental (values) decision, alternate composite scores should provide meaningful predictions, given that the intended purpose of setting an English-language proficient performance standard is to support decisions regarding sufficiency of English language proficiency with respect to the likelihood of attaining academic content proficiency. When alternate composites R^2 values drop to half that of the overall composite, comparisons between scores lose their meaning. The intention of an alternate composite score is to provide a *comparable metric* between ELs with disabilities and those without. Alternate composites with low comparative R^2 values have insufficient predictability and hence should not be used. With the aforementioned criteria, two scores have been identified as potential English-language proficient performance standard values for each alternate composite model: 318 and 323 for the literacy composite and 348 and 351 for the oral composite. Which score should be selected? A conservative approach is suggested here, i.e., select the highest score for each alternate composite model. For the literacy composite, a score of 323 would those be used to represent English proficiency. For the oral composite score, a value of 351 will be used to represent English proficiency.

Recommendation Activities in Creating an Alternate Composite

Since other experts might reasonably come to different conclusions, it is recommended that expert EL stakeholder groups be part of a decision-making process to establish criteria and "rules of thumb" for alternate composite scores. Below is a recommended sequence of activities for establishing alternate composite scores.

1. Determine potential alternate composite score combinations and provide rationale for the creation of each alternate composite score. This should be done with expert stakeholder input.

2. Conduct logistic regression analyses between content assessments and identified alternate composite scores at all tested grades. This will mean that grades where content assessments are not given will not have results.

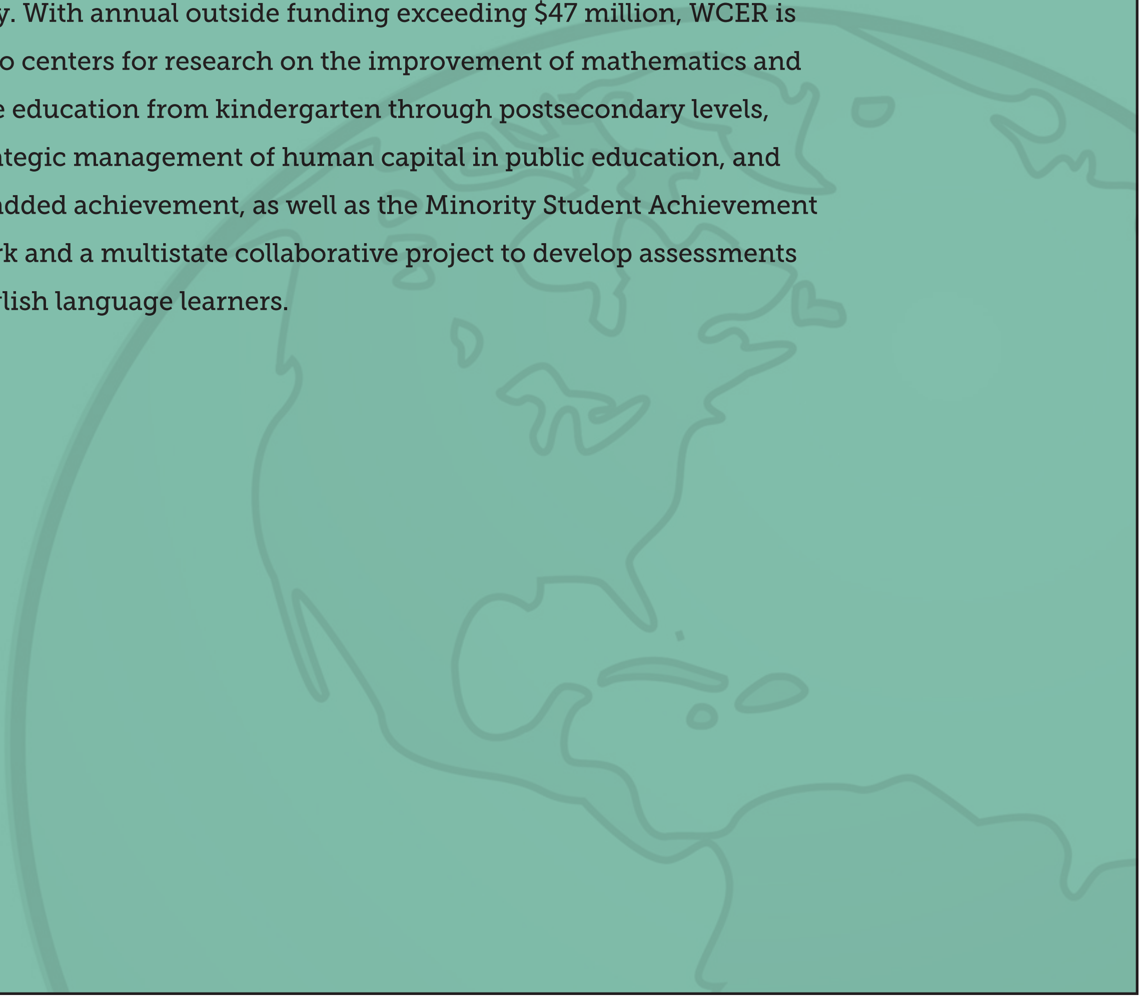
3. Convene an expert stakeholder group to determine acceptable goodness of fit values and alternate composite score probabilities that represent the English-language proficient performance standard for selected alternate composite models. Also, have the stakeholder group recommend alternate composite scores representing English-language proficiency in grades where a state content assessment is not administered.

4. Apply accepted alternate composite score proficient cut points to relevant AMAO formulae.

WCER

WISCONSIN CENTER FOR EDUCATION RESEARCH

The Wisconsin Center for Education Research (WCER) is one of the nation’s oldest university-based education research and development centers. WCER is based in the UW–Madison School of Education, which is consistently ranked one of the top schools of education in the country. With annual outside funding exceeding $47 million, WCER is home to centers for research on the improvement of mathematics and science education from kindergarten through postsecondary levels, the strategic management of human capital in public education, and value-added achievement, as well as the Minority Student Achievement Network and a multistate collaborative project to develop assessments for English language learners.



WIDA™ RESEARCH REPORT

Wisconsin Center for Education Research | University of Wisconsin-Madison
1025 West Johnson St. | Madison WI 53706 | www.wida.us

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON