Language Assessment Quarterly

ISSN: 1543-4303 (Print) 1543-4311 (Online) Journal homepage: http://www.tandfonline.com/loi/hlaq20

Exploring Domain-General and Domain-Specific Linguistic Knowledge in the Assessment of Academic English Language Proficiency

Anja Römhild , Dorry Kenyon & David MacGregor

To cite this article: Anja Römhild , Dorry Kenyon & David MacGregor (2011) Exploring Domain-General and Domain-Specific Linguistic Knowledge in the Assessment of Academic English Language Proficiency, Language Assessment Quarterly, 8:3, 213-228, DOI: 10.1080/15434303.2011.558146

To link to this article: https://doi.org/10.1080/15434303.2011.558146

Language Assessment Quarterly, 8: 213–228, 2011

ISSN: 1543-4303 print / 1543-4311 online
DOI: 10.1080/15434303.2011.558146

Routledge
Taylor & Francis Group

ARTICLES

Exploring Domain-General and Domain-Specific Linguistic Knowledge in the Assessment of Academic English Language Proficiency

Anja Römhild
University of Nebraska-Lincoln

Dorry Kenyon and David MacGregor
Center for Applied Linguistics, Washington, DC

This study examined the role of domain-general and domain-specific linguistic knowledge in the assessment of academic English language proficiency using a latent variable modeling approach. The goal of the study was to examine if modeling of domain-specific variance results in improved model fit and well-defined latent factors. Analyses were carried out on data from the ACCESS for ELLs® test battery, which comprises multiple test forms targeting different grade and proficiency levels. The results of the study provide empirical evidence in support of the conceptual distinction of domain-specific and domain-general linguistic knowledge. Domain-specific factors tended to become more salient with increasing language proficiency, whereas the salience of domain-general factors tended to decrease. However, overall domain-general factors remained stronger than the domain-specific factors. In one test form targeting high levels of proficiency, this factor pattern was reversed, suggesting some degree of fluidity in the relationship between domain-general and domain-specific linguistic knowledge.

Over the past two decades, academic English language (AEL) proficiency has become a major focus in the assessment of English language learners (ELLs) in the United States. Interest in academic English had been fueled by the dramatic increase in the number of ELL students in U.S. schools since the early 1990s. This development has brought many of the issues concerning the education of language minority students to the forefront. In particular, there has been a

Correspondence should be sent to Anja Römhild, University of Nebraska-Lincoln, 21 Teachers College, Lincoln, NE 68588-0353. E-mail: anja.roemhild@gmail.com

growing awareness among researchers and educators that ELL students, despite often adequate conversational English skills, tend to be poorly equipped for the language demands of classroom learning (Bailey & Butler, 2003; Scarcella, 2003; Solomon & Rhodes, 1995). In addition, ELL students tend to trail behind their English-speaking peers in academic performance (Abedi, Leon, & Mirocha, 2000/2005; Abedi, Lord, & Hofstetter, 1998; Butler & Castellon-Wellington, 2000/2005). Last, because of the limited English language skills of ELL students, concerns have been raised regarding the fairness and validity of test scores from mandatory academic content assessments (Abedi, 2004; Butler, Stevens, & Castellon, 2007).

In response to these issues, educational state agencies in the United States are placing increasing emphasis on the instruction and assessment of AEL in primary and secondary schools. This development has also been supported by the No Child Left Behind Act of 2001, which in addition to mandating annual assessments of ELL students' English language proficiency and academic achievement also requires states to align their English language proficiency standards with academic content standards. This latter provision has been instrumental in focusing English language instruction and assessment on language skills that are relevant to classroom learning and that enable students to meet academic content expectations. As a result, most U.S. states have now implemented English language assessments that focus at least in part on academic language skills (for a comprehensive overview of these assessments, see Porter & Vega, 2007).

Despite the attention given to academic language as an instructional goal and as a target for language proficiency assessments, the construct of AEL remains fairly obscure, particularly in the context of primary and secondary education. At those levels, academic language refers to the language of school and is broadly defined as "the language that is used by teachers and students for the purpose of acquiring new knowledge and skills" (Chamot & O'Malley, 1994, p. 40). Beyond this general definition, however, attempts to describe AEL in greater detail have reflected diverse viewpoints that have sought to differentiate academic language from other language varieties in terms of specific language functions, registers, levels of cognitive difficulty, or language use contexts (e.g., Cummins, 1980; Kinsella, 1997; Scarcella, 2003; Solomon & Rhodes, 1995).

Recent attempts to characterize the linguistic features of AEL have focused on the language used in specific academic content areas taught in school, such as English language arts, mathematics, science, and social studies (Gottlieb, 2004). Researchers at the National Center for Research on Evaluation, Standards, and Student Testing at the University of California, Los Angeles have conducted a number of empirical studies in which they identified specific language functions and lexical, grammatical, and discourse features of AEL across various fifth- and seventh-grade instructional and assessment materials as well as in classroom observations (Bailey, Butler, LaFramenta, & Ong, 2001/2004; Butler, Bailey, Stevens, Huang, & Lord, 2004; Butler, Lord, Stevens, Borrego, & Bailey, 2003/2004; Stevens, Butler, & Castellon-Wellington, 2000; Wolf et al., 2008). A general finding from this research is that linguistic features of AEL can be differentiated into those that are common to various academic content domains (i.e., domain-general features) and those that are unique to a specific content-domain (i.e., domain-specific features). The distinction between domain-general and domain-specific language features can also be applied to domains such as grade levels, as Bailey and Butler (2003) have done in their evidence-based research framework, which they developed to guide efforts to operationalize the AEL construct.

The distinction of domain-general and domain-specific academic language is conceptually intuitive, and the descriptive linguistic analyses just mentioned provide empirical evidence in

support of it. A question that has not been sufficiently addressed in the research literature concerns the role of domain-general and domain-specific linguistic knowledge in the acquisition process and in assessments of academic English. In particular, it is of interest to examine how both forms of linguistic knowledge contribute to proficiency in academic English and how that relationship may change with language development. For example, it is plausible that domain-general language features play a more important role in early academic language development, because these language forms are encountered more frequently in speech. Domain-specific language features, on the other hand, may be more likely to become prominent at later stages of development.

This study addresses the aforementioned research questions by examining how domain-general and domain-specific linguistic knowledge influence performance of ELL students on an AEL proficiency assessment at different levels of language development. Our study uses a confirmatory factor analysis approach to model domain-specific and domain-general variance and to evaluate and compare the salience of these variance sources against each other. The analyses are carried out on data from multiple test forms of the ACCESS for ELLs® English language proficiency test battery, which targets AEL proficiency at different grade clusters and proficiency levels (Bauman, Boals, Cranley, Gottlieb, & Kenyon, 2007). Conducting these analyses across multiple test forms allows us to compare the latent factor models across ELL student populations of different proficiency and grade levels.

DATA SOURCES

The ACCESS for ELLs English language proficiency test battery was developed by the World-Class Instructional Design and Assessment (WIDA) consortium and operationalizes the five English Language Proficiency Standards adopted by the WIDA consortium. The five WIDA standards describe performance expectations for ELLs in five specific contexts of language acquisition: social and instructional language, the language of language arts, the language of mathematics, the language of science, and the language of social studies (Gottlieb, Cranley, & Cammilleri, 2007). Each WIDA standard is assessed within the four language skills of listening, speaking, reading, and writing and at five proficiency levels. The overall test battery comprises a 12 test forms covering kindergarten through high school.[1] Test forms are grouped into grade-level clusters, which include early elementary grades (K–2), late elementary grades (3–5), middle school grades (6–8), and high school grades (9–12). Within each grade level cluster, there are three overlapping test forms that target levels of low (Levels 1–3), mid (Levels 2–4), and high (Levels 3–5) English language proficiency.

For this study, we analyzed data from the initial 2005 administration of the ACCESS for ELLs (Series 100), which was administered in the states of Alabama, Maine, and Vermont. Only data from the late elementary, middle school, and high school clusters were considered. Between 620 and 2,092 ELL students per test form participated in this test administration. A summary of the student populations is provided in Table 1.

[1]Since the first publication of the WIDA standards in 2004, there has been a revision of the Standards in 2007, which resulted in the expansion of the assessment framework to include a separate preschool (pre-K–K) grade-level cluster and an additional proficiency level targeting the upper end of the language proficiency continuum (Gottlieb et al., 2007).

TABLE 1
ELL Students' Ethnic Composition by Grade Cluster and Proficiency Level

Grade	*PL*	*N*	*Asian/ PI %*	*Black/ Non-H. %*	*Hispanic %*	*Am. Ind/Al %*	*Multiracial %*	*White/ Non-H. %*
3–5	Low	1,133	12.5	8.2	71.5	0.0	1.4	6.5
	Mid	2,092	12.5	6.1	71.2	1.5	1.4	7.3
	High	2,020	26.8	7.2	52.4	1.0	1.9	10.8
6–8	Low	854	10.3	7.1	73.0	0.2	1.7	7.8
	Mid	1,281	15.1	8.0	64.8	1.6	1.4	9.0
	High	1,646	26.1	4.2	54.5	1.5	2.5	11.2
9–12	Low	620	15.3	14.8	57.5	0.2	2.4	9.8
	Mid	1,058	23.3	11.8	51.2	0.2	3.0	10.5
	High	1,157	27.1	11.4	44.4	0.3	2.7	14.2

Note. Grade = Grade cluster; PL = Proficiency level; Asian/PI = Asian and Pacific Islander; Black/Non-H. = Black, Non-Hispanic; Am. Ind/Al = American Indian, Alaskan; White/Non-H. = White, Non-Hispanic.

METHOD

For each of the nine test forms considered in this study, we estimated various latent factor models that represent different configurations of domain-general and domain-specific linguistic knowledge. Analyses were conducted at the item level and include only items from the reading and listening subtests, since these subtests contain items assessing English proficiency in all five WIDA standards, that is, social and instructional language and language pertaining to the four academic subject areas language arts, mathematics, science, and social studies. To model domain-general linguistic knowledge, items from the reading subtest were specified to load on a "reading" factor and items from the listening subtest were specified to load on a "listening" factor with both skill factors allowed to correlate with each other. Domain-general linguistic knowledge is therefore represented as common variance pertaining to a specific language skill but which is nonspecific to any of the five language areas of the WIDA standards. Domain-specific linguistic knowledge, on the other hand, is represented by latent factors that are defined by items from both reading and listening subtests but which assess only one of the WIDA standards. As a result, in the final factor model each item loads on exactly one domain-general and one domain-specific factor. A simplified graphic of the final factor model with all five domain-specific factors and the two domain-general factors is provided in Figure 1 and described in further detail next.

Estimation of the factor models was carried out using Mplus 5 with a robust weighted least squares estimator (WLSMV). This estimator has been developed for analyses based on dichotomous outcome variables and has been shown to perform adequately with moderate sample sizes (Flora & Curran, 2004). Modeling began with a two-factor baseline model that includes only the reading and listening factors. This model corresponds to the factor structure generally assumed for skill-based English language proficiency assessments and does not assume domain-specific sources of variance. To examine the presence of domain-specific linguistic knowledge, we added latent factors representing each of the WIDA standards to the baseline model. These domain-specific factors were first added individually to test model convergence and improvement in model fit using chi-square difference tests. A final model combining all viable domain-specific

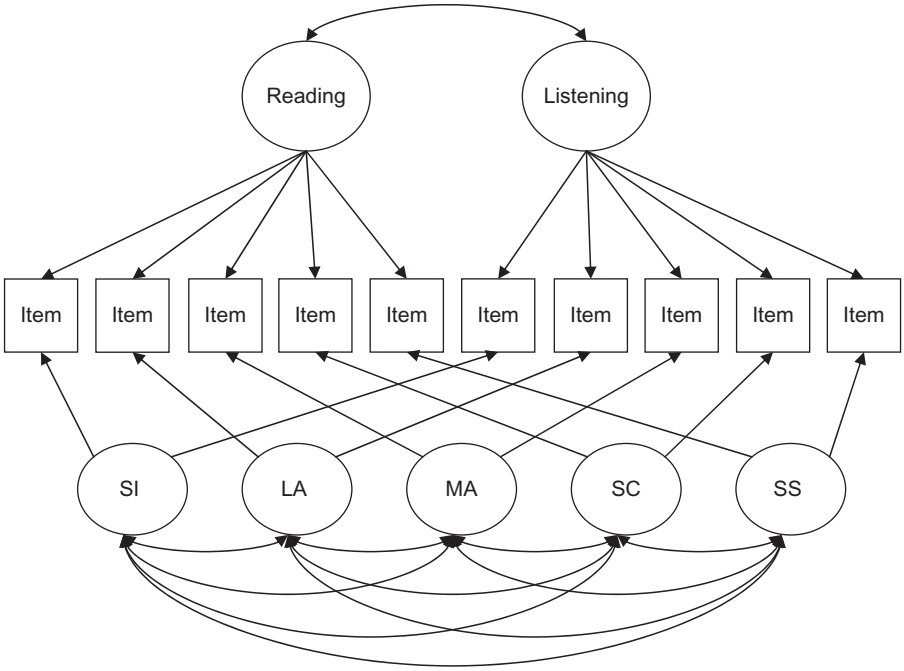

FIGURE 1 Latent factor model with domain-general and domain-specific language factors. *Note.* SI = social and instructional language; LA = language of language arts; MA = language of mathematics; SC = language of science; SS = language of social studies.

factors and the two domain-general factors was estimated, and model fit indices were compared to the baseline model to evaluate the impact of modeling domain-specific variance on model fit. In addition to model fit indices, we also examined the magnitude of the factor loadings in the final models, which permitted a closer look at the patterns of factor loadings and their changes within each factor and across. The factor loading patterns provide further information on the relative salience and interpretability of the domain-general and domain-specific factors across the different test forms.

To achieve convergence of the model solution, it was necessary to make model modifications in some instances. These modifications involved removing an ill-behaving item in the baseline and intermediate three-factor models and, in addition, imposing zero-correlations between the domain-specific factors in the full models. When modifications were made in the full model, preference was given to the latter strategy in order to keep as many items as possible.

Figure 1 presents a simplified depiction of a full factor model that includes all possible domain-specific and domain-general latent factors. Note that only two item indicators are represented for each WIDA standard (i.e., for each domain-specific factor). The actual test forms include between six and 16 items per WIDA standard spread between the reading and listening subtests with a minimum of three items per subtest. Accordingly, each box labeled "Item" in

Figure 1 represents a set of three items or more. The factor loadings of each latent factor are freely estimated, and latent factor variances are constrained to unity. In addition, the factors within the pair of domain-general factors and within the group of domain-specific factors are assumed to be intercorrelated, but no correlation was specified between these groups.[2] For the final model, the following model fit criteria were considered: comparative fit index values above 0.95, root mean square error of approximation values below 0.06, and weighted root mean square residual values below 1 (following recommendations in Yu, 2002). For model fit to be indicated, criteria of two of these fit indices had to be met.

RESULTS

The initial estimation of the two-factor baseline model yielded converged model solutions for six of the nine test forms. For the remaining three test forms, model convergence was obtained after the removal of one item.[3] Table 2 presents the model fit indices for each two-factor baseline model. As is shown, only the low-proficiency test forms in each grade cluster meet the fit criteria adopted in this study, which suggests that in these test forms a factor structure based on a conventional skills-based proficiency model provides an adequate representation of the latent test structure. If parsimony were desired in the selection of an appropriate latent variable model, then further specification of domain-specific factors for these test forms would not be necessary.

TABLE 2
Model Fit Indices for Two-Factor Baseline Model

	CFI	*RMSEA*	*WRMR*
Elementary school			
Low	**0.951**	**0.029**	1.133
Mid[a]	0.862	**0.026**	1.338
High[a]	0.905	**0.023**	1.220
Middle school			
Low	**0.950**	**0.024**	1.018
Mid[a]	0.892	**0.023**	1.124
High	0.905	**0.019**	1.085
High school			
Low	**0.952**	**0.024**	**0.970**
Mid	0.892	**0.026**	1.091
High	0.906	**0.023**	1.056

Note. Bold figures meet fit index criteria: comparative fit index (CFI) > .95, root mean square error of approximation (RMSEA) < .06, weighted root mean square residual (WRMR) < 1.0.

[a]One item was removed to achieve model convergence.

[2]In the intermediate factor models with a single domain-specific factor, the same zero correlation with the two domain-general factors was imposed.

[3]Reading Item 14 was removed in the mid- and high-level test form of the elementary grade cluster, and Listening Item 16 was removed in the midlevel test form of the middle school grade cluster.

TABLE 3
Statistical Significance ($\alpha < .05$) of Individual Domain-Specific Factors

	SI	*LA*	*MA*	*SC*	*SS*
Elementary school					
Low	sig.	sig.	sig.	sig.	sig.
Mid	sig.	sig.	sig.	sig.	*ns*
High	sig.	sig.	sig.	sig.	*ns*
Middle school					
Low	sig.	sig.	sig.	sig.	sig.
Mid	sig.	*ns*	sig.	sig.	sig.
High	sig.	sig.	sig.	sig.	sig.
High school					
Low	*ns*	*ns*	sig.	sig.	*ns*
Mid	sig.	sig.	sig.	*ns*	sig.
High	*ns*	sig.	sig.	sig.	sig.

Note. Statistical significance is based on mean and variance adjusted chi-square difference tests implemented in Mplus 5 with weighted least squares estimator estimation (Asparohouv & Muthen, 2006). SI = social and instructional language; LA = language of language arts; MA = language of mathematics; SC = language of science; SS = language of social studies.

After obtaining the two-factor baseline model solutions, we added the individual domain-specific factors to the model. Converged model solutions were obtained for all 45 three-factor models, which were then compared to the two-factor baseline model using mean and variance adjusted chi-square difference tests required in WLSMV estimation (Asparouhov & Muthen, 2006). The results are summarized in Table 3. The inclusion of the individual domain-specific factors resulted in statistically significant improvement in model fit in the majority of cases. Nonsignificant results were obtained in all but one of the WIDA standards, and there was only one test form, with multiple nonsignificant factors. In general, these results suggest that besides domain-general linguistic knowledge additional domain-specific dimensions are captured by the test items.

Given that all domain-specific factors were estimable, we proceeded to include all five factors in the final model for each test form. The full model allows us to evaluate the overall impact on model fit when all factors are included. In addition, we examined the patterns of factor loadings and their change in magnitude across test forms in order to evaluate the salience and interpretability of each latent factor.

To obtain converged model solutions, one test form had an item removed.[4] presents the model fit indices for the full factor models. As is shown, all test forms meet the fit criteria of at least two indices within rounding. On the left side of Table 4, we also provide the difference in the fit index values between the full model and the two-factor model. The differences in the comparative fit index and weighted root mean square residual are relatively large for most test forms. In general, there is improvement in model fit across all three fit indices and for all test forms including those that already meet the fit criteria under the two-factor baseline model. However, model fit improvement is particularly pronounced in the mid- and high-proficiency test forms, where the

[4]Listening Item 19 was removed in the low-proficiency test form of the Middle School grade cluster.

TABLE 4
Model Fit and Model Fit Improvement in Final Factor Models

	Full Factor Model			*Fit Index Difference*		
	CFI	*RMSEA*	*WRMR*	*ΔCFI*	*ΔRMSEA*	*ΔWRMR*
Elementary school						
Low	**0.964**	**0.024**	**0.981**	0.013	−0.005	−0.152
Mid[a]	0.945	**0.017**	1.036	0.083	−0.009	−0.302
High[a]	**0.967**	**0.013**	**0.964**	0.062	−0.01	−0.256
Middle school						
Low[a]	**0.967**	**0.02**	**0.924**	0.017	−0.004	−0.094
Mid[a]	0.944	**0.017**	**0.976**	0.052	−0.006	−0.148
High	**0.963**	**0.012**	**0.908**	0.058	−0.007	−0.177
High school						
Low	**0.963**	**0.021**	**0.909**	0.011	−0.003	−0.061
Mid[a]	0.949	**0.018**	**0.926**	0.057	−0.008	−0.165
High	**0.973**	**0.012**	**0.869**	0.067	−0.011	−0.187

Note. Bold figures meet fit index criteria: comparative fit index (CFI) > .95, root mean square error of approximation (RMSEA) < .06, weighted root mean square residual (WRMR) < 1.0.

[a]One item was removed to achieve model convergence.

two-factor model had failed to meet the fit criteria before. Overall these findings support the conclusion that domain-general variance and domain-specific variance are present and account for the underlying test structure in each test form.

A summary of the individual factor loadings is provided separately for each grade cluster in Tables 5 through 7. The tables provide information on the mean loading of each latent factor and the number of factor loadings that fall within certain ranges of magnitude. The ranges were chosen to reflect magnitudes that are considered trivial (below 0.2), weak (between 0.2 and 0.5), or moderate-high (above 0.5) by convention. It is assumed that loadings at or above 0.5 are sufficiently large to be considered salient. This value is slightly more conservative than suggested values reported in Tabachnick and Fidell (2001) and in Brown (2006). The factor loadings are summarized for each individual latent factor and for the groups of domain-general and domain-specific factors.

When the factor loading tallies are considered for each factor group, several patterns are discernable. In the low-proficiency test forms, the majority of the factor loadings for the domain-general factors are in the moderate-high category, whereas the factor loadings for the domain-specific factors are mostly in the trivial and weak range. This pattern suggests that the domain-general factors are fairly well defined but the domain-specific factors are not. The same conclusion can be drawn from the high mean factor loadings of the Listening and Reading factors, which are all above 0.5, in one instance as high as 0.7. In contrast, the mean factor loadings of the domain-specific factors and their group mean are either below or just around 0.2. These findings are consistent with the model fit results of the two factor baseline model in the low-proficiency test forms. In other words, when low levels of English proficiency are assessed, the amount of variance accounted for by the domain-specific factors is insubstantial and the addition of these factors plays a negligible role in explaining examinee performance on the test items.

TABLE 5
Summary of Factor Loadings for Low-, Mid-, and High-Proficiency Test Forms in Elementary School Grade Cluster

Elementary School	*Listening*	*Reading*	*SI*	*LA*	*MA*	*SC*	*SS*	*General*	*Specific*
Low proficiency									
No. of items	22	30	13	10	9	12	8	52	
<.2	0	1	10	8	5	6	5	1	34
.2–.5	4	9	3	2	4	6	2	13	17
>.5	18	20	0	0	0	0	1	38	1
M	0.70	0.57	0.18	0.14	0.17	0.20	0.21	0.63	0.18
Mid proficiency									
No. of items	21	29	8	16	9	11	6	50	
<.2	2	4	5	6	4	2	2	6	19
.2–.5	14	17	3	7	5	8	4	31	27
>.5	5	8	0	3	0	1	0	13	4
M	0.44	0.40	0.22	0.32	0.23	0.30	0.19	0.41	0.27
High proficiency									
No. of items	22	29	7	13	13	11	7	51	
<.2	2	2	5	5	9	7	4	4	30
.2–.5	13	13	1	5	4	3	3	26	16
>.5	7	14	1	3	0	1	0	21	5
M	0.42	0.48	0.19	0.30	0.15	0.21	0.14	0.46	0.20

Note. Factor loadings are based on absolute values. SI = social and instructional language; LA = language of language arts; MA = language of mathematics; SC = language of science; SS = language of social studies.

TABLE 6
Summary of Factor Loadings for Low-, Mid-, and High-Proficiency Test Forms in Middle School Grade Cluster

Middle School	*Listening*	*Reading*	*SI*	*LA*	*MA*	*SC*	*SS*	*General*	*Specific*
Low proficiency									
No. of items	20	26	11	10	13	6	6	46	
<.2	0	2	7	7	11	5	5	2	35
.2–.5	7	10	4	3	2	0	0	17	9
>.5	13	14	0	0	0	1	1	27	2
M	0.56	0.51	0.15	0.20	0.10	0.22	0.19	0.53	0.16
Mid proficiency									
No. of items	17	26	12	9	11	5	6	43	
<.2	2	1	6	8	9	2	2	3	27
.2–.5	8	16	6	1	2	3	3	24	15
>.5	7	9	0	0	0	0	1	16	1
M	0.43	0.46	0.22	0.13	0.12	0.18	0.26	0.45	0.17
High proficiency									
No. of items	22	25	6	14	12	9	6	47	
<.2	5	21	1	2	1	0	0	26	4
.2–.5	13	4	3	11	6	4	2	17	26
>.5	4	0	2	1	5	5	4	4	17
M	0.33	0.10	0.40	0.36	0.47	0.51	0.54	0.21	0.44

Note. Factor loadings are based on absolute values. SI = social and instructional language; LA = language of language arts; MA = language of mathematics; SC = language of science; SS = language of social studies.

TABLE 7
Summary of Factor Loadings for Low, Mid, and High Proficiency Test Forms in High School Grade Cluster

High School	*Listening*	*Reading*	*SI*	*LA*	*MA*	*SC*	*SS*	*General*	*Specific*
Low proficiency									
No. of items	22	28	11	9	9	14	7	50	
<.2	1	0	9	5	5	10	4	1	33
.2–.5	5	8	1	4	3	4	3	13	15
>.5	16	20	1	0	1	0	0	36	2
M	0.55	0.56	0.13	0.19	0.23	0.15	0.19	0.56	0.17
Mid proficiency									
No. of items	20	25	7	8	15	7	8	45	
<.2	0	0	4	5	5	4	5	0	23
.2–.5	10	13	3	3	8	3	3	23	20
>.5	10	12	0	0	2	0	0	22	2
M	0.56	0.48	0.23	0.17	0.28	0.23	0.15	0.51	0.22
High proficiency									
No. of items	21	29	6	12	12	13	7	50	
<.2	1	1	2	4	2	6	3	2	17
.2–.5	13	11	3	6	8	7	2	24	26
>.5	7	17	1	2	2	0	2	24	7
M	0.42	0.52	0.27	0.32	0.37	0.22	0.30	0.47	0.30

Note. Factor loadings are based on absolute values. SI = social and instructional language; LA = language of language arts; MA = language of mathematics; SC = language of science; SS = language of social studies.

In contrast, in the mid- and high-proficiency test forms, somewhat different patterns emerge. In general, one can see that the number of factor loadings in each range category has shifted in favor of the middle category that represents weak factor loadings. The increase in the number of weak loadings is accompanied by a decrease in loadings above 0.5 for the domain-general factors and a decrease in loadings below 0.2 for the domain-specific factors. This suggests that in the mid- and high-proficiency test forms, domain-specific factors appear to be somewhat more salient and correspondingly domain-general factors appear to be somewhat less salient than in the low-proficiency test forms.

This trend is also observable in the mean factor loadings for both factor groups. We plotted the means of the domain-specific and domain-general factor loadings for each proficiency level by grade cluster (Figures 2 and 3) and carried out Tukey-Kramer pairwise comparisons between the low-, mid-, and high-proficiency level test forms of each grade cluster to examine the statistical significance of the mean differences (Table 8) The graphs in Figures 2 and 3 show that with increasing proficiency level the means of the domain-general loadings tend to decline, whereas the means of the domain-specific loadings tend to increase. In addition, the pairwise comparisons resulted in at least one statistically significant mean difference between proficiency levels in all three grade clusters and for both domain-general and domain-specific factor loadings. These results suggest that the observed changes in the factor loadings are of meaningful magnitude.

Of particular note is the high-proficiency test form of the middle school cluster. In this test form, the mean factor loading of the domain-specific items surpasses the mean of the domain-general items. In addition, the majority of the domain-general factor loadings are in the weak or trivial range. The results for this test form suggest that under certain test conditions, performance

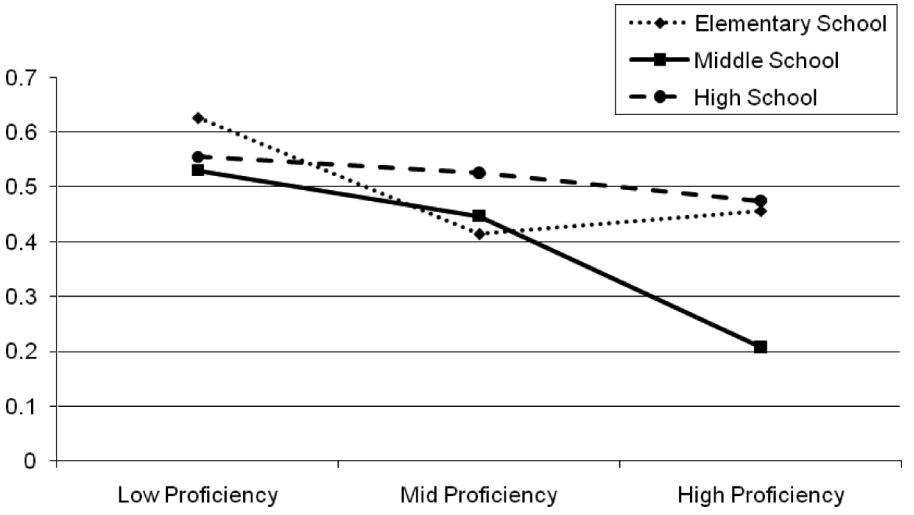

FIGURE 2 Means of domain-general item factor loadings.

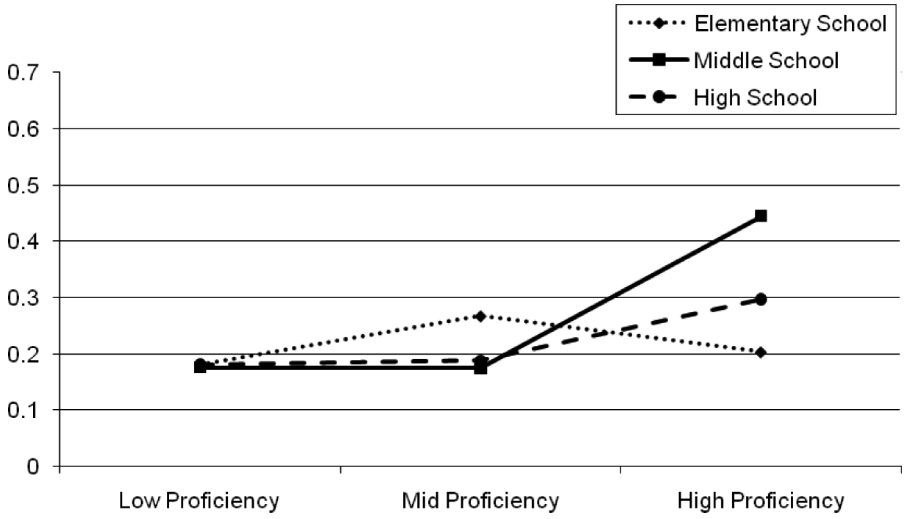

FIGURE 3 Means of domain-specific item factor loadings.

on test items can be more strongly driven by domain-specific factors than by domain-general factors. However, it is important to note that this is the only test form of nine where the salience of the domain-specific factor loadings surpassed that of the domain-general factor loadings. In all other test forms, the domain-general loadings clearly remain above the level of the domain-specific factor loadings.

Finally, when mean loadings of individual factors are considered within each factor group, we observed no systematic patterns that show a particular factor exhibiting consistently higher or lower loadings across all test forms. We computed Tukey-Kramer pairwise comparisons between

TABLE 8
Tukey-Kramer Pairwise Comparisons of Item Factor Loadings by Proficiency Level (PL)

	Elementary School		*Middle School*		*High School*	
	M Diff.	*p*	*M Diff.*	*p*	*M Diff.*	*p*
Domain-general						
PL1 vs. PL2	**0.21**	**<.001**	0.08	.058	0.03	.572
PL1 vs. PL3	**0.17**	**<.001**	**0.32**	**<.001**	**0.08**	**.013**
PL2 vs. PL3	−0.04	.468	**0.23**	**<.001**	0.05	.174
Domain-specific						
PL1 vs. PL2	**−0.09**	**.013**	0.0	.999	−0.01	.984
PL1 vs. PL3	−0.02	.735	**−0.27**	**<.001**	**−0.12**	**.003**
PL2 vs. PL3	0.06	.091	**−0.27**	**<.001**	**−0.11**	**.006**

Note. Bold figures are statistically significant. Mean difference based on absolute values of factor loadings.

the five domain-specific factors and independent *t* tests for the two domain-general factors to test whether any mean differences are statistically significant. The pairwise comparisons indicated no statistically significant mean difference in loadings between the five domain-specific factors in any of the nine test forms. There was a statistically significant mean difference between the reading and listening factors in the high-proficiency test forms of the middle school, $t(1) = 5.201$, $p < .001$, and high school, $t(1) = 2.968$, $p = .005$, cluster, but not in the other test forms. The two significant mean differences favored a different factor each and, therefore, provide no conclusive evidence. In general, the results of the mean difference tests suggest that the patterns and changes in factor loading magnitudes observed for the two factor groups are not the result of individual factors but seem to apply equally across all the factors within each group.

DISCUSSION

The purpose of this study was to investigate the relationship between domain-general and domain-specific linguistic knowledge in the assessment of AEL proficiency as operationalized in the WIDA ACCESS for ELLs test battery. To this end, we used a confirmatory factor analysis approach to model domain-general and domain-specific linguistic knowledge in a series of latent factor models. We first investigated the viability of individual domain-specific factors and the overall improvement in model fit they afford. We then examined the salience of the latent factors as groups by looking at the magnitude of factor loadings and compared them across different grade and proficiency levels.

The findings of our study generally support the conceptual distinction between domain-general and domain-specific linguistic knowledge. We were able to show empirically that both types of knowledge influence examinee's test performance on an English language proficiency test that measures social and academic language. We found that domain-specific latent variance, although not consistently present in all test forms, appears to influence examinee test performance when mid- or high levels of proficiency are assessed. Modeling the domain-specific

factors in these test forms generally improved overall model fit and produced higher, at times statistically significantly higher, factor loadings than in the low-proficiency test forms.

Despite this trend, however, we observed that the domain-general factors on the whole remained more salient and better defined than the domain-specific factors, which indicates that an assessment such as the ACCESS for ELLs test battery is primarily a measure of domain-general aspects of academic English proficiency and only secondarily a measure of domain-specific academic language knowledge. There can be exceptions, though, as the high-proficiency test form in the middle school cluster has shown. The observed relationship between the salience of the domain-specific factors and the proficiency level targeted by the grade cluster test forms suggests a process of AEL development, where aspects of domain-specific linguistic knowledge tend to be acquired at later stages of AEL development. On the other hand, domain-general aspects appear to play an important role throughout language development, but especially during early phases.

The presence of domain-specific latent variance in our study raises an interesting question regarding the relationship between academic language proficiency and academic content knowledge. Test developers are generally quick to emphasize that tests of academic English proficiency are not tests of academic content knowledge. However, it is difficult to differentiate between the two forms of knowledge when language specific to a certain content domain is assessed. For example, the ability to correctly produce and comprehend specialized academic terminology may require, at a minimum, a rudimentary understanding of the corresponding academic content matter. Given that domain-specific linguistic knowledge may play an increasingly more important role at mid- and high levels of proficiency, the distinction between academic language proficiency and academic content knowledge could also become increasingly blurred at these levels.

The influence of domain-specific linguistic knowledge on test performance also has implications for the development of AEL proficiency tests, in particular, for the construction of the score scale of the test. Although it is feasible that in many instances scaling models remain unaffected by the presence of secondary, domain-specific variance due to the high salience of the domain-general factors, in those cases where domain-specific variance represents the primary test dimension, the test-dimensional assumptions underlying the scaling model may be violated. Thus, it is crucial to investigate the presence of both variance sources, especially in test instruments targeting high levels of proficiency.

A related issue concerns the lack of stability in the dimensional test structure due to the strengthening of the domain-specific factors with increasing proficiency. In the ACCESS for ELLs, individual test forms target specific proficiency ranges, which allowed us to isolate the effect of domain-specific variance at these proficiency levels. Most language proficiency instruments, however, assess the entire proficiency continuum within a single test form with the underlying assumption that the test functions equivalently at all proficiency levels. Given the findings in this study, this assumption may not be easily met. Whether or not tests of academic English proficiency should measure the construct in the same dimensional configuration throughout their score scale requires a theoretical debate about the role of domain-specific linguistic knowledge within an academic language proficiency model. Such a debate has not taken place yet but would be crucial in order to achieve a clearer understanding of the psychometric validity of assessments of AEL proficiency.

CONCLUSIONS

The results of our study provide empirical support for the theoretical concepts of domain-specific and domain-general linguistic knowledge. Using a confirmatory factor analysis approach, we were able to show that both forms of knowledge can be represented as latent dimensions in an AEL proficiency assessment. Across the nine test forms in this study, we found domain-general linguistic knowledge to represent the primary dimensions in almost all cases and with domain-specific linguistic knowledge representing secondary test dimensions. However, although this was the predominant factor pattern in this study, we also observed in one test form that domain-specific linguistic knowledge can be the primary dimension, which suggests some fluidity in the relationship between domain-general and domain-specific linguistic knowledge. An interesting finding in this study concerns the increase in the salience of the domain-specific factors as a function of increasing proficiency. Although this relationship requires further replication, it is of interest in terms of what it suggests about the development of academic language proficiency and the acquisition of domain-specific and domain-general linguistic knowledge, in particular.

One of the strengths of the current study is the manner in which replications were performed not only across English language proficiency levels but also across three different grade clusters. Nevertheless, this study represents one of the first attempts to empirically examine AEL proficiency using a latent variable modeling framework, and more research is needed to gain a better understanding of the generalizability of the study's findings. In particular, the relationship between level of proficiency targeted by the assessment and the salience of domain-specific factors should be replicated on similar assessments. Future research should also consider the link between specific test characteristics and the occurrence of dominant domain-general or domain-specific factors, for example, by examining the linguistic characteristics of the test instruments exhibiting such factor patterns.

Finally, there is great need for research that focuses on the psychometric implementation of conceptual models of academic language proficiency. The dimensional complexity of language proficiency assessments often poses difficulties with conventional item response models, but sophisticated modeling solutions are available for complex multidimensional data. Yet they are rarely tested on language proficiency instruments. Research in this direction not only can help others to better understand the nature of academic language proficiency but also may produce more appropriate tools for the construction of valid and accurate assessments.

REFERENCES

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, *33*, 4–14.

Abedi, J., Leon, S., & Mirocha, J. (2000/2005). Examining ELL and non-ELL student performance differences and their relationship to background factors: Continued analyses of extant data. In *The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation From Three Perspectives* (CSE Tech. Rep. No. 663). Los Angeles: University of California, Los Angeles, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Lord, C., & Hofstetter, C. (1998). *Impact of selected background variables on students' NAEP math performance* (CSE Tech. Rep. No. 478). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Asparouhov, T., & Muthen, B. (2006). *Robust chi-square difference testing with mean and variance adjusted test statistics* (Mplus Web Notes: No. 10). Retrieved from http://www.statmodel.com/download/webnotes/webnote10.pdf

Bailey, A., & Butler, F. (2003). *An evidentiary framework for operationalizing academic language for broad application to K-12 education: A design document* (CSE Rep. No. 611). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Bailey, A., Butler, F., LaFramenta, C., & Ong, C. (2001/2004). *Towards the characterization of academic English in upper elementary science classrooms* (CSE Tech. Rep. No. 621). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Bauman, J., Boals, T., Cranley, E, Gottlieb, M., & Kenyon, D. (2007). Assessing comprehension and communication in English state to state for English language learners (ACCESS for ELLs®). In J. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 81 – 91). Davis: University of California, Davis, College of Education.

Brown, T. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford.

Butler, F., Bailey, A., Stevens, R., Huang, B., & Lord, C. (2004). *Academic English in fifth-grade mathematics, science, and social studies textbooks* (CRESST Tech. Rep. No. 642). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Butler, F., & Castellon-Wellington, M. (2000/2005). Students' concurrent performance on tests of English language proficiency and academic achievement. In J. Abedi, A. Bailey, F. Butler, M. Castellon-Wellington, S. Leon, & J. Mirocha (Eds.), *The validity of administering large-scale content assessments to English language learners: An investigation from three perspectives* (CSE Tech. Rep. No. 663). Los Angeles: University of California, Los Angeles, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Butler, F., Lord, C., Stevens, R., Borrego, M., & Bailey, A. (2003/2004). *An approach to operationalizing academic language for language test development purposes: Evidence from fifth-grade science and math* (CSE Tech. Rep. No. 626). Los Angeles: University of California, Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing.

Butler, F., Stevens, R., & Castellon, M. (2007). ELLs and standardized assessments: The interaction between language proficiency and performance on standardized tests. In A. Bailey (Ed.), *The language demands of school. Putting academic English to the test* (pp. 1–26). New Haven, CT: Yale University Press.

Chamot, A. U. & O'Malley, J. M. (1994). *The CALLA handbook: Implementing the cognitive academic language learning approach*. Reading, MA: Addison-Wesley.

Cummins, J. (1980). The construct of proficiency in bilingual education. In J. E. Alatis (Ed.), *Georgetown University round table on language and linguistics* (pp. 81–103). Washington, DC: Georgetown University Press.

Flora, D. B, & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*, 466–491.

Gottlieb, M. (2004). Overview. In *WIDA consortium K-12 English language proficiency standards for English language learners: Frameworks for large-scale state and classroom assessment. Overview document*. Madison: State of Wisconsin.

Gottlieb, M., Cranley, M. E., & Cammilleri, A. (2007). *Understanding the WIDA English Language Proficiency Standards. A resource guide*. World-class Instructional Design and Assessment Consortium. Madison, Wisconsin: Board of Regents of the Wisconsin System.

Kinsella, K. (1997). Moving from comprehensible input to "learning to learn" in content-based instruction. In M. A. Snow & D. M. Britton (Eds.), *Perspectives on integrating language and content* (pp. 46–68). White Plains, NY: Addison-Wesley Longman.

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

Porter, S. G., & Vega, J. (2007). Overview of existing English language proficiency tests. In J. Abedi (Ed.), *English language proficiency assessment in the nation: Current status and future practice* (pp. 93–103). Davis: University of California.

Scarcella, R. (2003). *Academic English: A conceptual framework* (University of California Linguistic Minority Research Institute Tech. Rep. No. 2003-1). University of California, Irvine.

Solomon, J., & Rhodes, N. (1995). *Conceptualizing academic language* (Research Rep. No. 15). Santa Cruz: University of California, National Center for Research on Cultural Diversity and Second Language Learning.

Stevens, R. A., Butler, F. A., & Castellon-Wellington, M. (2000). Academic language and content assessment: Measuring the progress of ELLs (CSE Tech. Rep. No. 552). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Tabachnick, B. G., & Fidell, L. (2001). *Using multivariate statistics* (4th ed.). Boston, MA: Allyn and Bacon.

Wolf, M., Herman, J., Kim, J., Abedi, J., Leon, S., Griffin, N., . . . Shin, H. (2008). Providing validity evidence to improve the assessment of English language learners (CRESST Rep. No. 738). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Unpublished doctoral dissertation, University of California, Los Angeles. Retrieved from http://www.statmodel.com/download/Yudissertation.pdf