University of Denver

Digital Commons @ DU

6-27-2022

High Quality, Equity, and Assessment: An Analysis of Variables Impacting English Learner Standardized Science Test Performance and Implications for Construct Validity

Maria del Carmen Salazar
University of Denver

Joanna K. Bruno
Colorado Department of Education

Melissa P. Schneider
Thompson School District

Follow this and additional works at: https://digitalcommons.du.edu/curr_inst_fac

Part of the Bilingual, Multilingual, and Multicultural Education Commons, Curriculum and Instruction Commons, Educational Assessment, Evaluation, and Research Commons, Language and Literacy Education Commons, and the Science and Mathematics Education Commons

Recommended Citation

High Quality, Equity, and Assessment: An Analysis of Variables Impacting English Learner Standardized Science Test Performance and Implications for Construct Validity

Article

High Quality, Equity, and Assessment: An Analysis of Variables Impacting English Learner Standardized Science Test Performance and Implications for Construct Validity

Maria del Carmen Salazar [1],* , **Joanna K. Bruno [2],*** and **Melissa P. Schneider [3],***

1   Morgridge College of Education, University of Denver, Denver, CO 80210, USA
2   Colorado Department of Education, Denver, CO 80203, USA
3   Thompson School District, Loveland, CO 80537, USA
*   Correspondence: maria.salazar@du.edu (M.d.C.S.); jotab12@yahoo.com (J.K.B.); melissapschneider@gmail.com (M.P.S.)

Abstract: In the United States, assessment is seen as a lever that can facilitate high-quality education. This study on English learners, students whose native language is not English, was based on data from eighth-grade English learners' performance on science general and content-specific (physical, life, and earth science) standardized exams and an English language proficiency exam. The researchers utilized regression analysis to examine factors (i.e., socioeconomic status, home language, English language proficiency, and receptive and productive elements of language) that are predictive of English learner performance in general and content-specific science standardized assessments to identify implications for construct validity of high-stakes science assessments. The research question is as follows: What factors influence the performance of English learners on a standardized science assessment, including overall performance and content-specific domains? Three main findings emerged from this research study. First, this study confirms previous research indicating that socioeconomic status and English language proficiency are predictive of English learners' achievement on content-based standardized tests. Second, this study adds to current research by providing evidence that productive language scores are the most significant predictors of English learner science achievement, in comparison to receptive language scores, overall English language proficiency scores, home language, and socioeconomic status. Third, this study adds to the body of evidence needed to challenge the validity of standardized science tests for English learner populations. The findings of this study challenge the construct validity of science content-based assessments for English learners, emphasizing the importance of productive language in academic performance.

Keywords: sustainable educational development; assessment; equity; science

1. Introduction

Educational assessment is a key aspect of high-quality education [1,2]. In the United States, assessment is seen as an essential lever to facilitate high-quality education. States across the U.S. mandate the use of standardized content tests to assess, monitor, and support English learner (EL) student learning. When standardized assessments are administered to ELs, questions of validity come into play due to a variety of factors. An EL is "an individual who has sufficient difficulty speaking, reading, writing, or understanding the English language to be denied the opportunity to learn successfully in classrooms where the language of instruction is English or to participate fully in the larger U.S. society" [3]. English learners can be found across the globe, but this article focuses on English learners in one state in the United States.

Standardized testing and questions of validity are exemplified in the state of Colorado in the U.S. In 2018, the percentage of ELs enrolled in public schools was 10% or higher in eight U.S. states, including Colorado [3]. In the 2016–2017 school year, ELs represented

approximately 12% of students in Colorado public schools; they were the fastest growing student population. The majority of ELs in Colorado, 77% and 80% respectively, are native Spanish speakers [4]. Colorado law requires every student to take standardized content assessments in English, regardless of the time enrolled in U.S. schools. The Colorado Measures of Academic Success (CMAS) is Colorado's standards-based assessment designed to measure the Colorado Academic Standards. The CMAS Science assesses four performance levels: distinguished, strong, moderate, and limited. The CMAS Science test results from 2015 point to a disparity between ELs and their non-EL peers. In 2015, statewide, 29% of non-ELs had distinguished and strong scores in the 8th grade CMAS Science test; however, only 6% of ELs scored at those two performance levels. Moreover, 61% of ELs received the lowest rating [5].

A significant body of research reveals that high-stakes assessments present a challenge for ELs [6–17]. Research points to three significant issues that impact the assessment of ELs. First, ELs' test performance may be negatively impacted due to high language demands in high-stakes assessments [6]. This assertion is supported by research that established a strong association between increasing linguistic complexity and decreasing test performance [7]. Second, a growing body of research demonstrates that ELs' test performance is reflective of their English language attainment versus their content knowledge [8–10]. As a case in point, in a study with 1700 ELs who were tested in English and Spanish on a standardized math achievement test, the results showed that the ELs answered more items correctly on a math test in their home language [11]. Third, researchers assert that assessment results for EL students are not valid—that is, the assessments are not measuring the intended construct [12,13]. For example, in the case of standardized science assessments, the intended construct is science content knowledge; however, when an assessment is given to ELs in English, the assessment measures their English language proficiency instead [14].

Researchers have established that there is a persistent disparity in test scores between ELs and non-ELs in academic content areas due to bias in testing [9,15]. However, there is a dearth of research on factors that impact EL test performance in general science *and* content-specific strands—and subsequently, the implications for the validity of high-stakes science tests [16]. It is vital to build a body of research that examines the extent to which current science testing practices adequately capture ELs' academic potential. The misuse of tests can lead to marginalization and discrimination toward immigrant and minority groups [17]. For instance, if testing practices do not adequately capture ELs' academic potential, this can lead to centering the "problem" of EL test performance on ELs themselves. This study builds on current research on persistent disparities in test scores by examining EL science test performance in general and content-specific strands based on factors known to impact EL student test performance (e.g., socioeconomic status, English language proficiency) and extending this research to less commonly examined factors (e.g., home language, productive and receptive elements of language).

The purpose of this study is to examine factors that are predictive of ELs' performance on general and content-specific science standardized tests and to identify implications for the construct validity of high-stakes science assessments. Specifically, we, the researchers collected and analyzed data from eighth-grade ELs' performance on the statewide CMAS Science general and content-specific (physical, life, and earth sciences) exams, as well as the English language proficiency exam—Assessing Comprehension and Communication in English State-to-State (ACCESS) for English Language Learners 2.0. We examined the following variables in relation to CMAS test performance: socioeconomic status, home language, English language proficiency, and receptive and productive elements of language. Socioeconomic status (SES) is defined as "the social standing or class of an individual or group" [18]. Home language (HL) is the native language spoken in the home [19]. In this research, home language variables were designated as HL Spanish and HL Other due to the fact that Spanish is the dominant language; approximately 80% of ELs speak Spanish in the state of Colorado. There is no consensus on the definitions of English language proficiency (ELP)—the definition depends on a variety of contextual factors [20]. Receptive language

and productive language are defined as the ability to comprehend language and the ability to produce language, respectively [21]. The research question is as follows: What factors influence the performance of ELs on a standardized science assessment, including overall performance and content-specific domains? The section that follows provides a synthesis of literature related to factors that impact test performance.

2. Literature Review

ELs' test performance is impacted by a combination of factors, including, but not limited to language proficiency, content knowledge, and sociocultural context. The following sections describe how these related variables inform the assessment of ELs' learning.

2.1. Language Proficiency

Students' language proficiency is typically examined within four domains of language: listening, speaking, reading, and writing [22]. Traditionally, these four domains have been divided into two sub-categories: oral language (listening and speaking) and literacy (reading and writing). Another way to combine the domains is "receptive" (listening and reading) and "productive" (speaking and writing) language. Receptive language refers to how well students receive and understand information, and productive language refers to how students produce and communicate that understanding. It is common to combine all four domains (either equally or weighted) to assess English language proficiency in high-stakes assessments or overall English language proficiency [23].

Academic language is essential in the development of language proficiency. Academic language is typically defined as a language register, or varieties of languages used for a particular purpose or setting, required for academic success [24]. Academic language can be categorized into general and discipline-specific language. General academic language is crosscutting language that includes vocabulary and structures found within many disciplines, whereas disciplines have their own academic discourse that includes content vocabulary and syntactic structures [24].

The "language of science" differs from other subject areas in that it includes specific discourse patterns, specialized semantic rules, and precise vocabulary [6]. The language of science is conveyed not just through oral or textual forms but also through visual and mathematical representations, including pictures, diagrams, graphs, charts, tables, maps, and equations [25]. Students need to master these nonlinguistic modes of representation to gain an understanding of science. Scientific language is often difficult for ELs to access due to the use of prepositional phrases, noun phrases, passive voice constructions, complex sentence structures, and an emphasis on high-level language skills such as argumentation and reasoning [26].

2.2. Content Knowledge

Researchers have long been interested in the relationship between content knowledge and language proficiency [10,27]. Science content and language are intertwined. For example, in science, students listen to and read, write, and speak about scientific concepts, and they visually represent scientific models. They also obtain, evaluate, and communicate their scientific understanding [28]. Moreover, each strand of science (physical, life, and earth) has different disciplinary discourse conventions. The differences are reflected in science assessments, which have language specific to a discipline and grade level. Students must absorb these differences as they work to construct meaning appropriate to the topic at hand. Each strand of science employs textual, nonlinguistic, or oral representations to different degrees. Thus, language is essential for demonstrating the content knowledge of science.

Understanding the connection between content and language has become increasingly relevant as large-scale assessments move toward an integrated approach of assessing content and language. In the content of science, language integration can include aspects such as verbal encoding of science concepts, grammatically encoded science language,

degree of meaning condensation, abstract and generalized linguistic forms, and robust academic vocabulary [15]. The integration of science content knowledge and language is complex, dynamic, and contextually dependent.

This complexity is captured in the Next Generation Science Standards (NGSS) that focus on three dimensions in learning science: disciplinary core ideas, science and engineering practices, and crosscutting concepts [29]. The new standards present a "deeper integration of science and language learning that . . . focuses on the need for all students, including ELLs to use language while engaging in science and engineering practice" [27] (pp. 397–398). As a result, the standards emphasize scientific discourse as a vehicle for understanding scientific ideas. State assessments across the U.S. have been aligned to the NGSS [30]; thus, the linguistic demands of state assessments are high for all students.

2.3. Sociocultural Context

Sociocultural factors can play a significant role in the language learning and content knowledge of ELs. Vygotsky's sociocultural theory describes a perspective on sociocultural factors that influence language development. Specifically, Vygotsky hypothesized that mental processes, such as language acquisition, take place through interaction with social and material environments [31]. ELs' language attainment and academic achievement are impacted by a variety of sociocultural factors, such as the initial language level, socioeconomic status (SES), parents' level of education, and access to challenging academic curriculum [24]. An examination of sociocultural factors is useful for interpreting assessment results in that students' performance on assessments may not only be a reflection of their cognitive abilities but also their social context.

2.4. Assessment of ELs

The development of valid assessments of EL achievement is an important scientific challenge [10,13]. Promising approaches to assessing ELs include multiple measures to assess the development of the desired result, while deficit approaches often use one measure to identify what is lacking in attaining the desired result [10]. State and district high-stakes assessment systems typically use a single measure to evaluate content learning and/or language deficits. Such systems are unfair and can in fact measure the wrong construct [10].

Assessment scholars have called for an approach to language and content assessment that moves beyond assertions of correct or incorrect performance, but rather toward an approach that provides a "more nuanced picture of learner abilities" [31] (p. 216). For example, current research on multilingual assessments demonstrates greater accuracy in assessing students' content knowledge because these assessments measure how students use language in authentic ways [10]. A growing body of research points to "the value of incorporating students' familiar communicative practices and experiences as supports for the development of new knowledge, whether scientific concepts or the discourses of science" [32] (p. 5). That is, research supports the inclusion of students' ways of knowing and communicating to build new science knowledge.

In contrast, high-stakes testing has resulted in a rigid, static, and narrow view of language correctness that ignores ELs' entire linguistic repertoire [32]. As a result, high-stakes assessments do not accurately capture what ELs know and are able to do; thus, they violate construct validity [12]. Construct validity is the accuracy of assessing a construct and the appropriateness of subsequent inferences and actions [10]. When an assessment is given in English and students have not yet attained English language proficiency, the assessment may not be measuring the content knowledge that it is intended to measure [14]. While research on assessment practices strongly discourages the use of single assessments to gauge students' language development and content learning, this is a common practice across states.

3. Method

This study examines the factors that influence the academic achievement of ELs in high-stakes science assessments, both general and content-based. The research question is as follows: What factors influence the performance of ELs on a standardized science assessment, including overall performance and content-specific domains? We, the researchers, used hierarchical multiple regression to conduct the analysis. This method allows for the testing of predictors in a particular order of theoretical interest to determine the predictive relationships between variables. Variables that are known to be predictive of variance, as stated in the hierarchical multiple regression literature, were added earlier to the model to account for a certain amount of the variance in achievement. Then, the variable(s) of interest were added later to the model to see if they are predictive of additional variance. Research has shown that SES and ELP are predictors of academic achievement for ELs. Therefore, these were included in the hierarchical regression to account for some of the variance. Home Language and Receptive and Productive elements were used in the model to see if there was any further explanation of variance above and beyond that of SES and ELP. We obtained the secondary data used for this analysis through CDE databases and included a Student Biographical Data Grid used in the two state-level assessments, CMAS Science and ACCESS for ELLs 2.0. The individual data were masked using a unique student identifier. Matched student identifiers were used in secondary data analysis.

3.1. Sample

The state of Colorado designates ELs into the following subgroups: NEP (Non-English Proficient), LEP (Limited English Proficient), FEP (Fluent English Proficient), FELL (Former English Language Learner), and PHLOTE (Primary Home Language Other Than English). NEP, LEP, and FEP are part of the Colorado Revised Statutes as official language designations for students who are learning English as a second language and who are receiving extra language support. FELL and PHLOTE are used for students who are not receiving extra support services, but whose language development is influenced by another home language other than English [4].

In 2015, 64,104 EL students took the CMAS Science assessment. The sample for this study was 6402 eighth-grade EL students who took the 2015 CMAS Science assessment and ACCESS for ELLs 2.0. Only the ELs coded as NEP and LEP who took both the CMAS Science and ACCESS exam were included in the sample. Per Colorado law, only ELs designated as NEP and LEP are required to take the ACCESS exam. Approximately 92% of the EL students who took the tests were identified as Spanish speakers.

3.2. Instruments

Colorado law requires that all students enrolled in public schools take the CMAS Science at the eighth-grade level, and students identified as NEP and LEP take an annual assessment of English language proficiency—ACCESS for ELLs 2.0. Further information on these assessments is provided below.

3.2.1. CMAS Science Tool

The CMAS is Colorado's standards-based assessment designed to measure the Colorado Academic Standards in science. For the 2015 CMAS, each assessment consisted of three sections. All sections contained a combination of selected-response items (28), technology-enhanced items (15), and constructed-response items (17). A subset of the science assessment includes simulation-based item sets, which are groups of items that relate to a scientific investigation or experiment. CMAS scores were validated using various sources of validity evidence, including the test content, response processes, internal structure, and fairness [5]. Reliability, using Cronbach's alpha, was reported at 0.93 for the overall assessment, and 0.82, 0.81, and 0.83 for the respective content domains: physical science, life science, and earth science [5]. This assessment design follows a universal design for a learning approach that specifically decreases the language load and removes

any extraneous language. Additionally, the integration of technology-enhanced items and simulations decrease the language load with regards to text.

3.2.2. ACCESS for ELLs 2.0 Tool

ACCESS is an English language proficiency test designed by the World-Class Instructional Design and Assessment (WIDA) consortium [33]. ACCESS for ELLs is the collective name for WIDA's suite of English language proficiency assessments. Colorado uses this instrument as their state English language proficiency assessment administered annually to ELs. ACCESS assesses academic language in language arts, mathematics, science, and social studies. Four composite scores are reported for the assessment: oral (listening and speaking domains), literacy (reading and writing domains), comprehension (listening and reading domains), and overall (listening, speaking, reading, and writing). In 2016, the ACCESS test for eighth-grade students had a composite reliability score of $\alpha = 0.930$. Evidence for the reliability and validity of the ACCESS exam is provided through the Center for Applied Linguistics (CAL) Validation Framework [33].

3.3. Analysis and Procedures

We, the researchers, were interested in variables that influence academic achievement of ELs in high-stakes science assessments, both general and content-based. This study includes both performance and demographic variables. The dependent variable was ELs' performance on the 8th grade CMAS Science assessment (overall scale score and scale score by content domain along a continuous scale). Four independent variables were socioeconomic status (SES), home language (HL), English language proficiency (ELP), and receptive and productive language (R&P). ELP and R&P are composite scores, and a stratified Cronbach's alpha coefficient was used to compute and weight the contribution of each domain score to determine the composite. SES was a control variable since research has established this as a predictor of achievement [34]. IBM SPSS 20 was used for all analyses in this research study.

The variables were entered as "blocks". This allowed for the testing of two models and the analysis of predictability of each individual variable. The first block displays SES (FRL—free and reduced lunch—and Non-FRL). The second block displays SES and primary home language (HL Spanish and HL Other). The third block displays SES, primary home language (HL), and overall English language proficiency (ELP), based on performance on the overall scale score. Block three includes receptive and productive elements of language. The receptive language "RL" and productive language "PL" are displayed as separate variables in combination. RL & PL are based on receptive (reading and listening) and productive (writing and speaking) domains of language. These variables were computed from the ACCESS data, specifically the "productive" composite variable. WIDA reports the "receptive" composite variable as comprehension, but does not calculate or report the "productive" composite variable. Therefore, we created this variable, using the same procedure as WIDA, by combining the speaking and writing domains of language based on the weights that WIDA used for each domain per their 2015 technical manual (speaking = 30% and writing = 70%) [33]. The calculation of the production score is consistent with the procedure WIDA used to calculate comprehension.

We used hierarchical multiple regression to evaluate the relationship between the independent variables and the dependent variable, controlling for the impact of a different set of independent variables on the dependent variable. Variables were entered into "blocks" in a fixed order of entry to control for the effects of covariates and to test the effects of certain predictors independent of the influence of others.

Multiple regression assumptions included linearity, homoscedasticity, independence of errors, and multicollinearity. The minimum sample size rule 5-to-1 was met; the sample was large. Analysis of residual plots revealed that the assumption of linearity was met. Outliers were examined using standardized residuals. The plot of these residuals was examined using the $+/-3$ rule to check for homoscedasticity. The Durbin–Watson test for

correlation of residuals was used to check for independence of errors using values between 1.5 and 2.5 [35]. Tolerance levels were investigated for multicollinearity for all independent variables, i.e., greater than 0.10. This assumption was not met. The section that follows reveals the results of the hierarchical multiple regression analysis.

4. Results

4.1. Predictors of Overall Science Achievement

We used hierarchical multiple regression analysis to examine overall achievement on the CMAS Science test as the criterion variable. Multiple regression assumptions for linearity and homoscedasticity were met. Multicollinearity was evaluated using a minimum tolerance level of 0.10 [35] and Variance Inflation Factor (VIF) maximum tolerance level of 10 [36]; the VIF recommendation of 10 corresponds to the tolerance recommendation of 0.10 (i.e., $1/0.10 = 10$). This assumption was violated for block four in the full regression model. The violation occurred between the overall ELP score and the R&P elements of the language scores. This is due to the R&P elements of language being inherently within the overall ELP. Therefore, to correct this violation, a three-block model used the original three blocks in the method as outlined previously, and then a second three-block model used the original two blocks as outlined, the overall ELP in block three was replaced with the R&P elements of language.

Table 1 displays the effect size measures (R^2), change in R^2, and adjusted R^2 for the full model, and Table 2 displays pooled unstandardized regression coefficients (*B*) and standardized regression coefficients (β). The changes in R^2 for each block suggest that for both models one and two, SES and primary home language combined accounted for 1.0% of the variability; then, by adding English language proficiency, model one accounted for 44% of the variability and model two, the receptive and productive elements of language, accounted for 48% of the variability in predicting the overall achievement regarding the CMAS Science test.

Table 1. Hierarchical Multiple Regression Model for Overall Achievement on the CMAS for Science Test.

Model	Block	R	R^2	ΔR^2	ΔF
	SES	0.09	0.01	0.01	47.49 ***
1	SES + HL	0.09	0.01	0.00	3.05
	SES + HL + ELP	0.67	0.44	0.44	5020.19 ***
	SES	0.09	0.01	0.01	47.49 ***
2	SES + HL	0.09	0.01	0.00	3.05
	SES + HL + R+P	0.69	0.48	0.48	2843.01 ***

Note. SES = Socioeconomic Status; HL = Home Language; ELP = English Language Proficiency; R & P = Receptive and Productive Elements of Language; *** $p < 0.001$.

Table 2. Hierarchical Multiple Regression Analysis Assessing HL, ELP, and R&P Elements of Language as Predictors of Overall Achievement on the CMAS for Science Test.

Model	Variable	Block One		Block Two		Block Three	
		B	B	B	B	B	B
	SES	−22.23 ***	−0.09 ***	−21.25 ***	−0.08 ***	−12.82 ***	−0.05 ***
1	HL			−6.1 ***	−0.02 ***	−10.97 ***	−0.04 ***
	ELP					2.76 ***	0.66 ***
	SES	−22.23 ***	−0.09 ***	−21.25 ***	−0.08 ***	−13.2 ***	−0.05 ***
2	HL			−6.1 ***	−0.02 ***	−8.54 ***	−0.03 ***
	RL					0.71 ***	0.13 ***
	PL					2.00 ***	0.59 ***

*** $p < 0.001$.

Table 2 displays two models. In the first model, block one, SES was a statistically significant predictor of academic achievement on the CMAS Science test, $F(1, 6400) = 47.49$, $p < 0.001$. In block two, SES and HL were statistically significant predictors of academic achievement on the CMAS Science test, $F(2, 6399) = 25.28$, $p < 0.001$. In block three, SES, HL, and ELP were statistically significant predictors of academic achievement on the CMAS Science test, $F(3, 6398) = 1703.47$, $p < 0.001$. These variables accounted for 44% of the variance in academic achievement on the CMAS Science test.

In the second model, block one, SES was a statistically significant predictor of academic achievement on the CMAS Science test, $F(1, 6400) = 47.49$, $p < 0.001$. In block two, SES and HL were statistically significant predictors of academic achievement on the CMAS Science test, $F(2, 6399) = 25.28$, $p < 0.001$. In block three, the SES, HL, and R&P elements of language were statistically significant predictors of academic achievement on the CMAS Science test, $F(3, 6398) = 1445.37$, $p < 0.001$. These variables accounted for 48% of the variance in academic achievement on the CMAS Science test. Therefore, the R&P elements of language increased the predictability of science achievement by an additional 4% over English language proficiency overall. It is important to note that the productive elements of language were more strongly predictive than the receptive language elements. All predictor variables had statistically significant correlations with overall CMAS Science achievement.

4.2. Predictors of Content Domains: Physical, Life, and Earth Science Achievement

Three different hierarchical multiple regressions were calculated to predict academic achievement in the three different strands of science (physical, life, and earth) within the CMAS Science test based on HL proficiency and R&P elements of language. For each strand, the first block displayed SES (FRL and Non-FRL). The second block displayed SES and HL (Spanish and Other). The third block displayed SES, HL, and overall ELP. An additional hierarchical regression was performed, including blocks one and two as stated above, and block three included the R&P elements of language as the predictor variables.

4.2.1. Physical Science

With the physical science strand on the CMAS Science test as the criterion variable, Table 3 displays the effect size measures (R^2), change in R^2, and adjusted R^2 for the full model, and Table 4 displays the pooled unstandardized regression coefficients (B) and standardized regression coefficients (β). The changes in R^2 in each block suggest that, for both models one and two, SES and HL combined accounted for 1.0% of the variability.

Table 3. Hierarchical Multiple Regression Model Summary for Physical Science Achievement on the CMAS for Science Test.

Model	Block	R	R^2	ΔR^2	ΔF
	SES	0.09	0.01	0.01	47.16 ***
1	SES + HL	0.09	0.01	0.00	6.55 **
	SES + HL + ELP	0.57	0.33	0.32	3052.28 ***
	SES	0.09	0.01	0.01	47.16 ***
2	SES + HL	0.09	0.01	0.00	6.55 **
	SES + HL + R+P	0.6	0.36	0.35	1741.92 ***

*** $p < 0.001$; ** $p < 0.01$.

Block one, SES, was a statistically significant predictor of physical science achievement on the CMAS Science test, $F(1, 6400) = 47.16$, $p < 0.001$. Block two (SES and HL) was a statistically significant predictor of academic achievement on the CMAS Science test, $F(2, 6399) = 26.88$, $p < 0.001$. Block three (SES, HL, and ELP) was a statistically significant predictor of academic achievement on the CMAS Science test, $F(3, 6398) = 1043.89$, $p < 0.001$. These variables accounted for 33% of the variance in academic achievement on the CMAS Science test. Block three (SES, HL, and R&P elements of language) was a statistically significant predictor of academic achievement on the CMAS Science test,

$F(3, 6398) = 891.71$, $p < 0.001$. These variables accounted for 36% of the variance in academic achievement on the CMAS Science test. Therefore, the R&P elements of language increased the predictability of the variability of achievement by an additional 3% over ELP overall, and the productive elements of language were more strongly predictive than the receptive elements.

Table 4. Hierarchical Multiple Regression Analysis Assessing Students' HL, ELP and R&P Elements of Language as Predictors of Physical Science Achievement on the CMAS for Science Test.

Model	Variable	Block One		Block Two		Block Three	
		B	B	B	B	B	B
1	SES	−24.3 ***	−0.09 ***	−22.73 ***	−0.08 ***	−14.80 ***	−0.05 ***
	HL			−9.8 **	−0.03 **	−14.38 ***	−0.05 ***
	ELP					2.59 ***	0.58 ***
2	SES	−24.3 ***	−0.09 ***	−22.73 ***	−0.08 ***	−15.32 ***	−0.05 ***
	HL			−9.8 **	−0.03 **	−11.64 ***	−0.04 ***
	RL					0.43 ***	0.07 ***
	PL					2.00 ***	0.54 ***

*** $p < 0.001$; ** $p < 0.01$.

4.2.2. Life Science

With the life science strand on the CMAS Science test as the criterion variable, Table 5 displays the effect size measures (R^2), change in R^2, and adjusted R^2 for the full models, and Table 6 displays the pooled unstandardized regression coefficients (B) and standardized regression coefficients (β). The changes in R^2 in each block suggest that in models one and two, SES and HL combined accounted for 1.1% of the variability.

Table 5. Hierarchical Multiple Regression Model Summary for Life Science Achievement on the CMAS Test for Science.

Model	Block	R	R^2	ΔR^2	ΔF
1	SES	0.07	0.01	0.01	32.78 ***
	SES + HL	0.08	0.01	0.00	5.79 *
	SES + HL + ELP	0.59	0.35	0.34	3330.43 ***
2	SES	0.07	0.01	0.01	32.78 ***
	SES + HL	0.08	0.01	0.00	5.79 *
	SES + HL + R+P	0.61	0.37	0.36	1834.61 ***

*** $p < 0.001$; * $p < 0.05$.

Table 6. Hierarchical Multiple Regression Analysis Assessing Students' HL, ELP, and R&P Elements of Language as Predictors of Life Science Achievement on the CMAS for Science Test.

Model	Variable	Block One		Block Two		Block Three	
		B	B	B	B	B	B
1	SES	−20.67 ***	−0.07 ***	−19.17 ***	−0.07 ***	−10.84 ***	−0.04 ***
	HL			−9.41 *	−0.03 *	−14.21 ***	−0.05 ***
	ELP					2.72 ***	0.58 ***
2	SES	−20.67 ***	−0.07 ***	−19.17 ***	−0.07 ***	−11.07 ***	−0.04 ***
	HL			−9.41 *	−0.03 *	−12.2 ***	−0.04 ***
	RL					0.89 ***	0.15 ***
	PL					1.88 ***	0.50 ***

*** $p < 0.001$; * $p < 0.05$.

Block one, SES, was a statistically significant predictor of physical science achievement on the CMAS Science test, $F(1, 6400) = 32.78$, $p < 0.001$. Block two (SES and HL) was a statistically significant predictor of academic achievement on the CMAS Science test,

$F_{(2, 6399)} = 19.30$, $p < 0.001$. Block three (SES, HL, and ELP) was a statistically significant predictor of academic achievement on the CMAS Science test, $F_{(3, 6398)} = 1129.70$, $p < 0.001$. These variables accounted for 35% of the variance in academic achievement on the CMAS Science test. Block three (SES, HL, and the R&P elements of language) was a statistically significant predictor of academic achievement on the CMAS Science test, $F_{(3, 6398)} = 932.48$, $p < 0.001$. These variables accounted for 37% of the variance in academic achievement on the CMAS Science test. Therefore, the R&P elements of language increased the predictability of the variability of life science achievement by an additional 2% over ELP overall, with the productive elements of language more strongly predictive than the receptive elements of language.

4.2.3. Earth Science

With the earth science strand on the CMAS Science test as the criterion variable, Table 7 displays the effect size measures (R^2), change in R^2, and adjusted R^2 for the full models, and Table 8 displays the pooled unstandardized regression coefficients (B) and standardized regression coefficients (β). The changes in R^2 in each block suggest that, in both models one and two, SES and HL combined accounted for 1.0% of the variability.

Table 7. Hierarchical Multiple Regression Model Summary for Earth Science Achievement on the CMAS for Science Test.

Model	Block	R	R^2	ΔR^2	ΔF
1	SES	0.07	0.01	0.01	33.02 ***
	SES + HL	0.07	0.01	0.00	0.22 ***
	SES + HL + ELP	0.57	0.33	0.32	3049.05 ***
2	SES	0.07	0.01	0.01	33.02 ***
	SES + HL	0.07	0.01	0.00	0.22 ***
	SES + HL + R+P	0.59	0.35	0.35	1714.88 ***

*** $p < 0.001$.

Table 8. Hierarchical Multiple Regression Analysis Assessing Students' HL, ELP, and R&P Elements of Language as Predictors of Earth Science Achievement on the CMAS for Science Test.

Model	Variable	Block One		Block Two		Block Three	
		B	B	B	B	B	B
1	SES	−20.94 ***	−0.07 ***	−20.64 ***	−0.07 ***	−12.48 ***	−0.04 ***
	HL			−1.85 *	−0.01 *	−6.56 *	−0.02 *
	ELP					2.67 ***	0.57 ***
2	SES	−20.94 ***	−0.07 ***	−20.64 ***	−0.07 ***	−12.94 ***	−0.04 ***
	HL			−1.85 *	−0.01 *	−3.94	−0.01
	RL					0.55 ***	0.09 ***
	PL					2.00 ***	0.53 ***

*** $p < 0.001$; * $p < 0.05$.

Block one, SES, was a statistically significant predictor of physical science achievement on the CMAS Science test, $F_{(1, 6400)} = 33.02$, $p < 0.001$. Block two (SES and HL) was a statistically significant predictor of academic achievement on the CMAS Science test, $F_{(2, 6399)} = 16.62$, $p < 0.001$. Block three (SES, HL, and ELP) was a statistically significant predictor of academic achievement on the CMAS Science test, $F_{(3, 6398)} = 1032.71$, $p < 0.001$. These variables accounted for 33% of the variance on the CMAS Science test. Block three (SES, HL, and the R&P elements of language) was a statistically significant predictor of academic achievement on the CMAS Science test, $F_{(3, 6398)} = 870.20$, $p < 0.001$. These variables accounted for 35% of the variance in academic achievement on the CMAS Science test. Therefore, the R&P elements of language increased the predictability of the variability of earth science achievement by an additional 2% over ELP overall, with

the productive elements of language being more strongly predictive than the receptive elements of language.

5. Discussion

This study addresses the need to further research the role that language plays in EL standardized science test performance—general and content-specific strands. Three main findings emerged from this research study. First, this study confirms previous research indicating that SES and ELP are predictive of ELs' achievement on content-based standardized tests [9,10,20]. This is the case in the content area of science overall and by content strand. Second, this study adds to current research by providing evidence that productive language scores are the most significant predictors of EL science achievement in comparison to receptive language scores, overall ELP scores, home language, and SES. Third, this study adds to the body of evidence needed to challenge the validity of standardized science tests for EL populations. The sections that follow further discuss these findings.

5.1. Factors that Impact EL Science Test Performance

There were key findings related to the impact of SES, HL, and ELP on EL standardized test achievement in science.

5.1.1. SES

Previous studies examining SES and student achievement demonstrated that SES is a statistically significant predictor of standardized test performance [37]. This conclusion has also been reached in research on the factors that impact EL achievement, that is, ELs with high SES outperform ELs with low SES on standardized assessments [8,38]. This study confirms that SES is an important predictor of academic achievement in standardized science testing. However, this study found that when combined with ELP and receptive and productive elements of language, the latter better predicts achievement than SES.

5.1.2. Home Language

There is scant research related to test performance comparison by home language. Generally, researchers consider linguistic variation within and across groups [39]. This study found home language to be a significant predictor of test performance. In the case of the data set used in this research, Spanish was the dominant language; approximately 80% of ELs speak Spanish in the state of Colorado [40]. It was not possible to disaggregate data based on language variety or country of origin based on the available state data. Further research is needed to disaggregate, understand, and explain the impact of home language on test performance. For example, distinctions in test performance by home language may be due to variations in SES, segregation, and a lack of opportunity to learn rigorous and applied science content.

5.1.3. English Language Proficiency

In both overall science and by strand, ELP was a statistically significant predictor of achievement. As EL students' ELP scores increase, their science achievement scores increase. This is consistent with prior studies looking at ELP as a predictor of achievement [41,42]. In this study, ELP was found to predict achievement more strongly than SES.

5.2. Receptive and Productive Language

Receptive and productive elements of language were statistically significant predictors of both overall science achievement and by strand compared with ELP alone. As EL students' receptive and productive scores increase, their science achievement scores increase. However, this study found that productive elements of language were the most significant predictors of achievement. Research supports the impact of productive language. In a study of 274 adolescent first-generation immigrant students from China, the Dominican Republic,

Haiti, Central America, and Mexico, the amount of time students were immersed in social situations where they spoke English was highly predictive of their ELP [43]. However, there is scant research on the impact of receptive and productive elements of language on standardized science test achievement. This issue warrants additional focus and research.

6. Conclusions

While this research is based on state-level data, it adds to the body of research on science assessment practices for ELs and challenges the construct validity of standardized content mastery tests. The findings of this study have implications that include challenging the construct validity of science content-based assessments for ELs and emphasizing the importance of productive language in academic performance. Moreover, it is important to address limitations of the research and delineate opportunities for future research.

6.1. Research Contribution

In the context of one state, when used to measure science knowledge of ELs, the CMAS also functions as a language assessment. This threatens construct validity and implies measurement variance among student groups [44]. In the U.S. national context, the Next Generation Science Standards (NGSS) amplify linguistic demands [27] for ELs; as state standardized science tests align with these standards, this has important implications for ELs. It is vital to build a research-based validity argument against the exclusive use of standardized tests to assess proficiency in science content knowledge and skills. This study found that content-specific science assessments are equally as problematic as general science assessments. Standardized-science assessments that are based on standards with high linguistic demands do not properly assess student content learning [14]. The result is construct-irrelevant variance or "scores are affected by a variable that is unrelated to that which the test is intended to measure" [9] (p. 96). This has implications for the validity, or accuracy, of standardized science assessments, both general and content-specific.

Policymakers need to take into consideration a students' level of ELP when requiring the assessment of ELs who are not proficient in the language of the assessment. Researchers concur that, "Some proportion of the academic achievement gap may be due not to an EL's lack of content knowledge, but to the content assessment's inability to accurately measure that knowledge when insufficient language proficiency stands in the way" [26] (p. 278). That is, the gaps in standardized test performance are likely due to the test's inability to measure content knowledge due to language proficiency. It is important to generate research that challenges the validity of science content-based standardized assessments for ELs. It is not enough to tinker with the tests; instead, standardized tests must be wholly reconceptualized. It is vital to develop accurate measurement tools that provide the necessary information to make appropriate inferences about the progress and needs of ELs. Such tools must include inclusive, democratic, and formative approaches "aiming to use tests in constructive and positive ways" [17] (p. 441). Moreover, it is essential to develop and research alternative methods of assessment, such as multilingual assessments that include, but are not limited to multilingual vocabulary, opportunity to use bilingual scaffolds, use of translated items, modified items to reduce linguistic complexity, and linguistically simplified tests [45].

6.2. Limitations and Future Research

First, one of the limitations of this study was that ELs were grouped together. It is known that this population of students is very diverse; however, the background characteristics of ELs were unavailable to the researchers and were thus incalculable. Second, in terms of home language, the grouping of languages other than Spanish was not predictive of achievement. It is possible that the "n size" accounts for a non-predictive result. When n sizes are mismatched or too small, then the inference made from the results could be difficult to determine. Third, the results rely heavily on the receptive and productive elements of language and due to their nature as composite scores, they are compensatory. Thus, "a

high score in one language domain could inflate the composite score, compensating for a low score in another language domain; conversely, a low score in a language domain could bring down the composite" [33] (p. 9). This means that caution is needed to interpret the composite scores.

Further research is needed on the role of productive language in standardized test achievement. This is perhaps the most impactful finding of this research study. The National Research Council proposes a dramatic rethinking of science education grounded in deemphasizing discrete facts with a focus on breadth over depth to one that provides students with engaging opportunities to experience how science is performed [46]. Mastery of the NGSS demands an evolved three-dimensional approach to teaching, learning, and assessing where science is seen not only as a noun, something to learn, but also as a verb, something to do. Inherent in this "doing" is productive language. In generating the language of science, students demonstrate verbal encoding of science concepts, grammatically encoded science language, degree of meaning condensation, abstract and generalized linguistic forms, and robust academic vocabulary [15].

The research on productive elements of language is almost exclusively focused on language proficiency and literacy development. More research is needed to explore the impact of productive elements of language in different content areas and the impact on standardized test achievement. This is particularly vital in science, as the NGSS have increased in linguistic complexity and thus increased the language demands of standardized science tests. Moreover, it is important to continue to probe the findings of such research by content-specific areas because the linguistic demands can vary.

Research on factors that impact ELs' science standardized test achievement is important in the short term in order to help ELs demonstrate academic success. Concomitantly, it is vital to look to the future and re-envision science assessment systems. The National Research Council states that with the onset of the NGSS, students will need multiple and varied assessment opportunities to demonstrate their understanding [47]. Building a body of evidence is standard practice when assessing EL students for their English proficiency, and that practice should be considered for science as well. The field needs an assessment system that is inclusive of summative, formative, and dynamic assessments meant to capture what EL students know and can do.

References

1. Xue, E.; Li, J.; Peters, M. High-Quality Education Equity for Promoting Educational Sustainable Development: Message from Guest Editors. Available online: https://www.mdpi.com/journal/sustainability/special_issues/high_quality_education_equity_promoting_educational_sustainable_development (accessed on 23 March 2022).
2. Dima, A.M.; Meghisan-Toma, G.M. Research on implementing education for sustainable development. In Proceedings of the 12th International Conference on Business Excellence, Bucharest, Romania, 22–23 March 2018; Volume 12, pp. 300–310.

3. National Center for Educational Statistics. English Language Learners in Public Schools. 2019. Available online: https://nces.ed.gov/programs/coe/indicator_cgf.asp (accessed on 23 March 2022).

4. Colorado Department of Education [CDE]. Every Student Succeeds Act (ESSA): The Colorado Plan Summary Report. 2017. Available online: https://www.cde.state.co.us/fedprograms/essasummary (accessed on 23 March 2022).

5. Colorado Department of Education [CDE]. CMAS Science Assessment Technical Manual. 2016. Available online: http://www.cde.state.co.us/assessment/newassess-sum (accessed on 23 March 2022).

6. Lee, O. Science education with English language learners: Synthesis and research agenda. *Rev. Educ. Res.* **2005**, *75*, 491–530. [CrossRef]

7. Wolf, M.; Leon, S. An investigation of the language demands in content assessments for English Language Learners. *Educ. Assess.* **2009**, *14*, 139–159. [CrossRef]

8. Abedi, J. *Impact of Student Language Background on Content-Based Performance: Analysis of Extant Data*; National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, University of California: Los Angeles, CA, USA, 2003.

9. Avenia-Tapper, B.; Llosa, L. Construct relevant or irrelevant? The role of linguistic complexity in the assessment of English language learners' science knowledge. *Educ. Assess.* **2015**, *20*, 95–111. [CrossRef]

10. Mahoney, K. *The Assessment of Emergent Bilinguals: Supporting ELLs*; Multilingual Matters: Clevedon, UK, 2017.

11. Abella, R.; Urrutia, J.; Shneyderman, A. An examination of the validity of English-language achievement test scores in an English language learner population. *Biling. Res. J.* **2005**, *29*, 127–144. [CrossRef]

12. Ary, D.; Cheser Jacobs, L.; Sorensen Irvine, C.K.; Walker, D.A. *Introduction to Research in Education*; Cengage Learning Inc.: Toronto, OA, Canada, 2019.

13. Solano-Flores, G.; Li, M. Generalizability theory and the fair and valid assessment of linguistic minorities. *Educ. Res. Eval.* **2013**, *19*, 245–263. [CrossRef]

14. Bailey, A.; Carroll, P. Assessment of English language learners in the era of new academic content standards. *Rev. Res. Educ.* **2015**, *39*, 253–294. [CrossRef]

15. Oliveira, A.W.; Weinburgh, M.H. *Science Teacher Preparation in Content-Based Second Language Acquisition*; Springer: Berlin/Heidelberg, Germany, 2017.

16. Baily, A. Language analysis of standardized achievement tests: Considerations in the assessment of English Language Learners. In *The Validity of Administering Large-Scale Content Assessments to English Language Learners: An Investigation from Three Perspectives*; Abedi, J., Bailey, A., Butler, F., Castellon-Wellington, M., Leon, S., Mirocha, J., Eds.; Center for the Study of Evaluation: Los Angeles, CA, USA, 2005; pp. 79–100.

17. Shohamy, E. Critical language testing. In *Language Testing and Assessment: Encyclopedia of Language and Education*; Shohamy, E.E., Or, I., May, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2017; pp. 441–454.

18. American Psychological Association. Socioeconomic Status. 2018. Available online: http://www.apa.org/topics/socioeconomic-status/ (accessed on 23 March 2022).

19. Polinsky, M.; Kagan, O. Heritage Languages: In the "Wild" and in the Classroom. *Lang. Linguist. Compass* **2017**, *1*, 368–395. [CrossRef]

20. Wolf, M.K.; Kao, J.; Herman, J.; Bachman, L.F.; Bailey, A.; Bachman, P.L.; Farnsworth, T.; Chang, S.M. Issues in assessing English language learners: English language proficiency measures and accommodation uses. *Pract. Rev. CRESST Rep.* **2008**, *732*, 1–75.

21. Rose Weinert, S.; Ebert, S. The roles of receptive and productive language in children's socioemotional development. *Soc. Dev.* **2018**, *27*, 777–792. [CrossRef]

22. Short, D.; Fitzsimmons, S. *Double the Work: Challenges and Solutions to Acquiring Language and Academic Literacy for Adolescent English Language Learners*; Alliance for Excellent Education: Washington, DC, USA, 2007.

23. Saville-Troike, M.; Barto, K. *Introducing Second Language Acquisition*; Cambridge University Press: Cambridge, UK, 2017.

24. Bailey, A.; Huang, B. Do current English language development/proficiency standards reflect the English needed for success in school? *Lang. Test.* **2011**, *28*, 343–365. [CrossRef]

25. Quinn, H.; Lee, O.; Valdes, G. *Language Demands and Opportunities in Relation to the Next Generation Science Standards for English Language Learners: What Teachers Need to Know*; Stanford University: Stanford, CA, USA, 2012.

26. Bailey, A. *The Language Demands of School: Putting Academic English to the Test*; Yale University Press: New Haven, CT, USA, 2007.

27. Llosa, L.; Lee, O.; Jiang, F.; Haas, A.; O'Connor, C.; Van Booven, C.D.; Kieffer, M.J. Impact of a large-scale science intervention focused on English language learners. *Am. Educ. Res. J.* **2016**, *53*, 395–424. [CrossRef]

28. Hakuta, K.; Santos, M.; Fang, Z. Challenges and opportunities for language learning in the context of the CCSS and the NGSS. *J. Adolesc. Adult Lit.* **2013**, *56*, 451–454. [CrossRef]

29. Next Generation Science Standards Lead States. *Next Generation Science Standards: For States, by States*; The National Academies Press: Washington, DC, USA, 2013.

30. Next Generation Science Standards. Criteria for Procuring and Evaluating High-Quality and Aligned Summative Science Assessments. 2018. Available online: https://www.nextgenscience.org/sites/default/files/Criteria03202018.pdf (accessed on 23 March 2022).

31. Lantolf, J.; Thorne, S.L.; Poehner, M. Sociocultural theory and second language development. In *Theories in Second Language Acquisition*; Van Patten, B., Williams, J., Eds.; Routledge: London, UK, 2015; pp. 207–226.

32. Poza, L.E. The language of *ciencia*: Translanguaging and learning in a bilingual science classroom. *Int. J. Biling. Educ. Biling.* **2018**, *21*, 1–19. [CrossRef]

33. World Class Instructional Design and Assessment [WIDA]. Annual Technical Report for ACCESS for ELLs. 2015. Available online: http://www.cde.state.co.us/assessment/ela-additionalresources (accessed on 23 March 2022).

34. Lee, O.; Luykx, A. Dilemmas in scaling up innovations in elementary science instruction with nonmainstream students. *Am. Educ. Res. J.* **2005**, *42*, 411–438. [CrossRef]

35. Tabachnick, B.G.; Fidell, L.S. *Using Multivariate Statistics*, 6th ed.; Allyn and Bacon: Boston, MA, USA, 2013.

36. Bankston, C.L.; Caldas, S.J. *Public Education—America's Civil Religion: A Social History*; Teachers College Press: New York, NY, USA, 2015.

37. Hair, J.F.; Black, W.C.; Babin, B.J.; Anderson, R.E.; Tatham, R.L. *Multivariate Data Analysis*; Pearson: Upper Saddle River, NJ, USA, 2006; Volume 6.

38. Krashen, S.; Brown, C.L. The ameliorating effects of high socioeconomic status: A secondary analysis. *Biling. Res. J.* **2005**, *29*, 185–196. [CrossRef]

39. Solano-Flores, G.; Li, M. Language variation and score variation in the testing of English language learners, native Spanish speakers. *Educ. Assess.* **2009**, *14*, 180–194. [CrossRef]

40. Colorado Department of Education [CDE]. Culturally and Linguistically Diverse Learners in Colorado: State of the State. 2016. Available online: http://www.cde.state.co.us/cde_english/stateofstate (accessed on 23 March 2022).

41. Abedi, J.; Gándara, P. Performance of English language learners as a sub-group in large-scale assessment: Interaction of research and Policy. *Educ. Meas. Issues Pract.* **2006**, *25*, 36–46. [CrossRef]

42. Solano-Flores, G.; Trumbull, E. Examining language in context: The need for new research and practice paradigms in the testing of English language learners. *Educ. Res.* **2003**, *32*, 3–13. [CrossRef]

43. Carhill, A.; Suárez-Orozco, C.; Páez, M. Explaining English language proficiency among adolescent immigrant students. *Am. Educ. Res. J.* **2008**, *45*, 1155–1179. [CrossRef]

44. Kline, R. *Principles and Practice of Structural Equation Modeling*, 4th ed.; Guilford Press: New York, NY, USA, 2016.

45. Abedi, J.; Lord, C. The language factor in mathematics tests. *Appl. Meas. Educ.* **2001**, *14*, 219–234. [CrossRef]

46. National Research Council. *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*; The National Academies Press: Washington, DC, USA, 2012.

47. National Research Council. *Developing Assessments for the Next Generation Science Standards*; The National Academies Press: Washington, DC, USA, 2014. [CrossRef]