R305A170151 Executive Summary
Improving the Accuracy of Academic Vocabulary Assessment for English Language Learners

Executive Summary of Year 2 Progress

In year 2 we continued our investigation of differential item functioning and its explication through explanatory item response models. We resolved the problem of estimation of DIF that had arisen in our study because of cohort differences in ability that complicated analyses due to confounding of items and cohorts. Specifically, some items were included only on the 2010-2011 test form, some items were included only on the 2011-2012 test form, and some items were included on both test forms. When DIF was examined using different methods, specifically multi-level generalized mixed linear models versus IRT models, we obtained highly concordant estimates of DIF for items on the 2010-2011 test form and items on the 2011-2012 test form that were also included on the 2010-2011 test form. However, estimates were very different for items that appeared only on the 2011-2012 test form and only for the Target Group (ELs). After considerable effort to resolve the discrepancy, we simulated the problem in an artificial dataset that precisely reproduced the problem that we were seeing in real data and confirmed that the IRT estimates were biased due to the difference in cohort ability between the 2011-2012 and 2010-2011 cohorts. This ability difference was due to the decision to limit the participants in the data analysis to first time test takers in 2011-2012, which meant that the 2011-2012 cohort was comprised of predominantly sixth grade students, along with small samples of students who were new arrivals to the schools in grades 7 and 8, whereas the 2010-2011 cohort was comprised of 6^{th}, 7^{th}, and 8^{th} grade students. This problem was important to rectify before progressing to descriptive and explanatory DIF models, otherwise we risked having biased estimates of DIF for some items. Because the generalized mixed linear models include an explicit measure of ability in the model, cohort differences apear to be controlled more effectively than they are in the IRT estimation approach in this context.

We continued the build out of the database of item and word characteristics and the development of tools for studying the effects of item characteristics on DIF and on item difficulty. In last year's report, we detailed the word level characteristics that were collected, cleaned, and merged. This year we focused on understanding the relationship between these characteristics and approaches to data reduction that could be applied to reduce an unwieldy number of variables for explanatory models. Our work has been informed by lexical processing studies that explore multiple word-level factors to explain speed and accuracy in lexical decision and naming tasks. We omitted subjective, behavioral measures (e.g. reaction time, accuracy) in our data reduction analysis, focusing on 19 discrete, objective variables (e.g. length, frequency; see Appendix 1 of the full report). We conducted an exploratory factor analysis (EFA) with an oblique rotation, identifying five factors (Orthographic Complexity, Proximity, Frequency, Semantic Diversity, Word Senses) for our academic words, which explained 60% of the variance in word features. This factor structure is similar to structures reported by other investigators. We are investigating the role that features play in item difficulty and in particular in differential item difficulty between EO's, and EL's, RFEP's, and IFEP's, although the latter group performs quite comparably to EO's.

We have also begun the process of identifying items for revision, and/or replacement, which we expect to complete over the summer along with the creation of new pilot items.

Executive Summary of Entire Project from Year 1

Purpose

The purpose of this project is twofold. First, we wish to improve understanding of factors that affect performance of English language learners (ELs) on the assessment of academic vocabulary knowledge. Second, based on these findings, we will refine an existing assessment to be psychometrically and theoretically sound for ELs. The aforementioned objectives will be completed in three phases. Phase I will focus on an archival data analysis, will not require data collection. The archival data analysis will utilize explanatory item response models to better understand factors that affect academic vocabulary test performance, and will help in refining the existing assessment. Phase II will involve writing new test items based on information acquired in Phase I, and conducting a pilot study to examine the refined assessment. Explanatory item response models, differential item functioning and differential distractor functioning analyses will be used to examine psychometric properties of the redesigned assessment. Phase III will concentrate on collecting data and validating the improved assessment. The project is currently in Phase I.

Project Activities

Project activities vary across the phases of the project. In Phase I, activities involve coding of words and items for specific features of interest and the construction of a large database that links information from various sources about words, their meanings, and tests/items. This data can then be compiled to create the Q-side design matrix for explanatory item response models for various vocabulary assessments, of which the Word Generation and Gates McGinitie vocabulary assessments are the two primary examples used in the current study. While our focus is on the current assessments, the design of the database takes into consideration extension to other assessments. In addition to the coding of word, meaning, and item features, we are analyzing data from a previously collected study to examine items that function differently for ELs and non-ELs in the 6-8th grades, in order to ascertain those items that function differently and the reasons that items function differently for ELs and non-ELs. In Phase II of the project, we will make use of the information from Phase I and the resultant database to construct new items for the Word Generation vocabulary assessment in order to improve the assessment of academic vocabulary for ELs. Phase II activiites include the construction of items and piloting them with small samples of students to ensure their general suitability for inclusion in a revised test. Finally, in Phase III, we will administer, score, and validate the new Word Generation assessment with a large sample of ELs and non-ELs across grades 6-8.

Products

In Phase I the primary products are the individual databases and the integrated database and any tools/guidance for working with them, as well as the results of analyses of DIF, DDF, and e-IRT models of our extant tests. Results will be disseminated largely through peer-reviewed publications and web reports. In Phase II, the products are new items and test forms for the Word Generation assessment of Academic Vocabulary. In Phase III, the products will include new test forms as well as peer-reviewed publications and web reports on the new assessment.

Study Design and Participants

In Phase I, the participants are middle school students from schools in northern California who participated in the Word Generation Efficacy study. Specifically, we are conducting archival analysis of the pre-test assessments from that efficacy trial. It is possible that we will also examine post-test assessments to address questions about consistency of models and inferences about item features, but the plan calls for archival analysis of data from the first occasion of meaurement for each participant. The sample size is approximately 13,780 in Phase I, 30-50 in Phase II (Piloting),and 1,050 in Phase III (Validation). Students in Phase I (N = 13,780) attended thirteen middle schools in a large urban district in California during the 2010-2012 academic years (Year 1: 2010-2011, Year 2: 2011-2012).

English proficiency status

The participating district identified language minority students in three categories: 1) initially fluent English proficient (IFEP; those who had gained full English proficiency by the time they enter school), 2) redesignated fluent English proficient (RFEP; those originally classified as limited English proficient who attained sufficient English proficiency to be reclassified), and 3) limited English proficient (LEP; those whose limited English proficiency continued to qualify them for language support or ELLs). Those who were not language minority learners were classified as English-only (EO) students.

Table 1 describes the number of students in each English proficiency status by grade levels in the beginning of each academic year. The collaborating district identified English proficient (i.e., IFEP, RFEP) and non-proficient (i.e., ELL) language minority students, and we will use this classification in our analysis to test whether English proficient language minority students (many of whom formerly used to be ELLs) and ELLs differ in their response to academic vocabulary test items. Note, this analysis does not simply compare ELLs and non-ELLs on the probability of answering a given item correctly, but assesses response comparability conditional on ability.

Table 1. Number of Students who Participated in the Data Collection by their Grade Levels and English Proficiency Status (total *N* = 13,780).

		Language Proficiency Status			
		EO	IFEP	RFEP	LEP
Year 1	6th	879	205	854	472
	7th	877	254	972	454
	8th	869	311	1022	449
	All	2,625	770	2,848	1,375
Year 2	6th	589	189	699	228
	7th	771	206	972	299
	8th	705	249	1,012	243
	All	2,065	644	2,683	770

Note. EO = English-only; IFEP = Initially fluent English proficient; RFEP = Redesignated fluent English proficient; LEP = Limited English Proficiency

In Phase III of this project we will test the new instrument using a new sample of students. At the time that our proposal was submitted to the funding agency, we had obtained the cooperation and support of a specific district in CA. However, given the departure of a Co-Investigator from CAdddf since the time of submission, we may elect to collect data in a different district to facilitate the data collection. This decision will be made closer in time to when data will be collected.

Measures and Analytic Strategy

The primary measures in Phase I are the Word Generation Assessment of Academic Vocabulary and the reading comprehension and vocabulary tests of the Gates McGinitie Reading Test. The analysis strategy in Phase I includes logistic regression and item response models for the assessment of Differential Item, Distractor, and Test functioning, and explanatory item response models (E-IRT) to identify item and person characteristics and their interactions that affect test functioning, with the specific goal in mind of isolating those features of items that affect the comparability of items for EL and non-EL students of similar ability. Phase II involves descriptive analyses, whereas Phase III involves pyschometric analyses to examine item and test functioning, test validity, and normative performance.

Three general approaches will be used in the data analyses: Explanatory Item Response Models, Differential Item Functioning, and Differential Distractor Functioning. In Phase III, we will also apply these methods to newly collected data to validate the performance of the new assessment.

Differential Item Functioning (DIF)

The question of differential item functioning is being addressed through logistic regression as well as through item response models.

Logistic Regression Approach. To detect whether vocabulary items function differently

for EOs and ELLs, we will conduct logistic regression analyses for each vocabulary item (Swaminathan and Rogers, 1990) using *SAS* (SAS Institute Inc., 2008). In the logistic regression models: (a) the outcome variable will be a correct response to an item on the academic vocabulary test, (b) the student's total test score on the academic vocabulary test will serve as the vocabulary knowledge ability score, and (c) the ELLs (i.e., RFEP and LEP) will represent focal groups and the EOs will serve as a reference group. We believe that including the IFEP as a group in these analyses will not be necessary. There is no reason to expect that the test will be different for IFEP's and EO's. For each item, the DIF analyses will be conducted in several steps to determine the presence of DIF, and the type of DIF (uniform vs nonuniform; Guler and Penfield, 2009). The analyses will be computed in two sets corresponding to the two ELL groups (i.e., RFEP and LEP). In each set of logistic regression analyses, the grouping variable will only include two categories: (a) EO vs RFEP, or (b) EO vs LEP. This approach as opposed to including multiple dummy variables representing different groups of English language proficiency levels in one model will be preferable because it will allow us to compute separate tests to detect DIF.

IRT Approach. DIF will be also investigated with the 1PL descriptive item response model (dIRT) using *IRTPRO 2.1* software (Cai, Thissen, & du Toit, 2011). The analyses will be computed in two sets corresponding to the two ELL groups (i.e., RFEP and LEP). DIF testing will consist of two steps, and will not require designated anchors. In the first step, we will fit a model wherein mean = 0 and variance = 1 for EO students, the mean and *variance* for ELL students are estimated, and item difficulty is constrained to be equal for EO and ELL students. In the second step, a model will be fitted with the ELL mean and *variance* fixed to the values obtained in the first step. This will link the metric between EO and ELL groups, and then all item parameters will be free to vary between groups. As a result, it will be possible to compute the Wald test for each item as a test for DIF.

Explanatory Item Response Models

Application of eIRT has become popular because of their focus on simultaneously explaining item responses on a test in terms of: (a) the effects of student characteristics on person ability (θ_p - one's location on a latent trait continuum), and (b) the effects of item features on item difficulty (β_i - difficulty of an item designed to measure some latent ability; De Boeck & Wilson, 2004). The goal of the eIRT models is to jointly explain a student's position on the vocabulary knowledge dimension as a function of student characteristics, and an item's position on the difficulty dimension as a function of item features. The eIRT models are advantageous when compared to standard psychometric models for tests or statistical models for ability because they: (a) take into account correlations among the items, (b) allow for estimation of individual differences, (c) allow for modeling of random and fixed effects of person and/or item parameters as they belong to a broader class of the generalized linear mixed models, (d) explain the probability of correct responses utilizing external variables, and (e) jointly model the probability of correct responses as a function of person and item characteristics. In the context of current project, application of these models is the most suitable because it allows for the joint modeling of student characteristics, target word characteristics, their interactions, and their effects on performance on vocabulary test items.

We propose to use the one-parameter item response model (1PL), reconfigured as an explanatory item response model as described by DeBoeck and Wilson, (2004), and von Davier,

Rost, and Carstensen (2007). In the 1PL model, the relation between item performance and ability, referred to as item discrimination, is constrained to be the same for all test items, whereas item difficulty is allowed to vary across items. Thus, items differ from one another only in terms of how difficult they are. Placing a constraint on the discriminability parameter carries important implications for interpretation of the unknown parameters and scoring of the test. Specifically, the restriction not only implies that the test is unidimensional and measures a single latent ability, but also implies that the number of correct item responses is a sufficient statistic for person ability, that is, there is a one-to-one mapping between the number correct and person ability. The 1PL model further implies that the probability of correctly answering a more difficult item can never exceed the probability of correctly answering an easier item for individuals of any given ability level. The same is not true for the 2PL and 3PL.

In the context of the current project the 1PL rather than the 2PL or 3PL models will be more desirable because: (1) our academic vocabulary test items are equally weighted, (2) preliminary analyses suggest that our test is unidimensional, (3) 1PL offer consistency and ease of scoring, and (4) our research questions focus on investigating factors affecting item difficulty rather than both item difficulty and item discrimination (Embretson & Reise, 2000).

Before computing the 1PL eIRT models we will examine the AIC and BIC fit indices of descriptive (i.e., without external variables) 1PL, 2PL, or 3PL models. The AIC and BIC indices will be used because there are better suited with measures including a large number of items (Cai, Thissen, & du Toit, 2011). Although fit indices will be examined, these statistics will not be used as a primary source of the model selection because model fit or misfit does not necessary prove whether an IRT model is useful or not for a given purpose. Specifically, (a) model that shows a substantial degree of misfit may still prove its usefulness in a given context (Kenny, 2015, Maydeu-Olivares, 2015), and (b) correlations between ability estimates of 1PL, 2PL, and 3PL models exceed .9.

In the computed 1PL eIRT models, we will use target word characteristics, students' characteristics (i.e., English language proficiency level, general vocabulary knowledge, and grade), and their interactions as the external variables explaining probability of correct responses to vocabulary test items. The interactions between the level of English language proficiency and target word characteristics will help to determine whether certain target word characteristics increase difficulty of test items for ELLs at different proficiency levels. Specifically, four classes of models will be estimated: (a) an unconditional model without explanatory variables, (b) Model 1 - model with target word characteristics, (c) Model 2 - model with students' characteristics, and (d) Model 3 - the final model – an interactive model that includes the explanatory variables separately investigated in models 1 and 2, and their two-way interactions.

Phase III Analyses

With the data collected in Phase III, we plan to conduct similar sets of analyses that were done in Phase I. We will do eIRT, DIF, and DDF analyses to see whether we were able to overcome the shortcomings of the previous version of our assessment. We will also examine internal consistency and validity of the assessment to confirm that our refined assessment is a reliable and valid test of Academic Vocabulary for EL and non-EL students. Lastly, we will look for items that conform to our design principles, but still function poorly, to test the applied validity of these principles for redesigning test items.

R305A170151 Annual Progress Report – Year 2
Improving the Accuracy of Academic Vocabulary Assessment
for English Language Learners

I. ACCOMPLISHMENTS

What are the major goals of this project?

The purpose of this project is to improve understanding of factors that affect performance of English language learners (ELs) on the assessment of academic vocabulary knowledge, and then, based on this improved understanding, to refine an existing assessment to be psychometrically and theoretically sound for ELs. The project is organized into three phases. Phase I focuses on an archival data analysis of a large data set from a previously completed intervention study that included a large sample of middle-school (i.e., grades 6-8) ELs and non-ELs and two different assessments of vocabulary and an assessment of reading comprehension. The archival data analysis will utilize logistic regression and item response models to examine differential item and distractor functioning (DIF and DDF, respectively), and will then use explanatory item response models to better understand factors that affect academic vocabulary test performance for all students and to understand specifically the factors that make items differentially difficult for ELs. This information will be crucial in guiding our plans to refine the existing assessment. Phase II will involve writing new test items based on information acquired in Phase I, and conducting a pilot study to examine the refined assessment. Explanatory item response models, differential item functioning (DIF) and differential distractor functioning (DDF) analyses will be used to examine psychometric properties of the redesigned assessment following completion of the revised assessment and the collection of a large validation sample in Phase III. Phase III will concentrate on collecting data and validating the improved assessment using explanatory item response models and analyses of DIF and DDF. The project is currently in Phase I.

What was accomplished with respect to these goals?

Project Year 2 Milestones and Progress (in bold font); Year 1 Milestones (not bolded); Year 2 Progress on Year 1 Milestones (in bold font)

Milestone	Accomplishment	Status
1. Fully staff the project	*Project is fully staffed for year 1 after hiring a post-doctoral Fellow, Dr. Autumn McIlraith, who completed her Ph.D. in Communication Sciences and Disorders from Florida State University under the direction of Dr. Hugh Catts. We have also recruited requisite non-key personnel to assist in item coding tasks.*	*100%*
2. Code item level	*We have completed coding of all items for many features (details provided below), but have identified*	***100%***

characteristics	*additional characteristics of words and meanings that we continue to code for both items and distractors. These additional features are anticipated to be useful in the crafting of new items and in understanding the functioning of existing items.* ***All item coding has been completed, including adding some new features.***	
3. ***Complete Item-Level Analyses.*** *Make substantial progress on item-level analyses (i.e., eIRT, DIF, DDF)*	*In year 1, we made substantial progress on item-level analyses, primarily in the analysis of Differential Item Functioning (DIF). We have completed logistic regression analyses of DIF using different approaches to the estimation of ability, viz., using an ability estimate from the test being studied, an ability estimate of the same ability using a test other than the studied test, and using a latent ability estimate from a factor analysis of items from both measures. These analyses have helped to identify items that systematically show DIF, which may be either uniform DIF or non-uniform DIF, as well as items whose characterization varies depending on the ability estimate. When using an estimate of ability that is an observed test score, more items are identified that show DIF that favors the target group, rather than the reference group. When a latent ability estimate is used, items showing DIF almost all show DIF that favors the reference group, which is more consistent with theoretical expectations. (Details provided below). We are beginning eIRT analyses and Differential Distractor Function (DDF) analysis.* ***In year 2, we made substantial progress on item analyses. We have completed DIF analyses, including reconciling discrepancies across estimation approaches (GLIMMIX and IRT). When DIF was examined using multi-level generalized mixed linear models versus IRT models, we obtained highly concordant estimates of DIF except for items that appeared only on the 2011-2012 test form. After considerable effort to resolve the discrepancy, we simulated the problem in an artificial dataset that precisely reproduced the problem that we were seeing in real data, confirming that the IRT estimates were biased due to the difference in cohort ability between the 2011-2012 and 2010-2011 cohorts. This problem was important to rectify before progressing to descriptive and explanatory DIF models. Because the generalized mixed linear models include an explicit measure of ability in the model, cohort differences are***	*65%/**90%***

	controlled more effectively than they are in the IRT estimation approach. Thus, our simulation work confirmed that the GLIMMIX estimates of DIF were to be preferred over the IRT estimates. Subsequent eIRT analyses have been undertaken descriptively by relating difficulty and DIF estimates to item characteristics for target words, key words, and distractor words. eIRT and DDF analyses are ongoing.	
4. Based on Year 1 analysis results, refine existing vocabulary assessment items;	***We have begun identifying the items to be revised/replaced, but have not yet rewritten or piloted revised items. We are waiting to complete DDF analyses and eIRT analyses to ensure that we develop the best possible new items***	***50%***
5. Iteratively pilot test the revised vocabulary assessment items with approximately 30 middle school students;	***We expect to pilot test items and revise them in the fall of project year 3 and then to administer the revised test in the spring of year 3.***	***0%***
6. Analyze data collected from the pilot test;	***We expect to analyze data from pilot test items in the fall of project year 3 and then to administer the revised test in the spring of year 3.***	***0%***
7. Refine and retest items as necessary;	***We expect to revise and refine the test in the fall of project year 3, and to obtain data on the revised test in the spring of year 3.***	***0%***
8. Prepare a project website; and	***We have begun development of the project website. We have secured several domain names for the project, and have begun the development of content focused on differential item performance, data on word and item characteristics, and user driven data displays that allow users to explore the relations in the already collected data. We have developed SHINY Apps in R for use on the website. We expect the website to be operational by the end of grant year 2 (August 31, 2019)***	***20%***
9. ***Submit any accepted peer-reviewed scholarly manuscripts to ERIC***	***No manuscripts have been published to date. Results of DIF and eIRT analyses were included in a symposium at AERA in Toronto.***	***100%***

Detailed Report on Year 2 Accomplishments

Code item level characteristics

In last year's report, we detailed the word level characteristics that were collected, cleaned, and merged. This year we focused on understanding the relationship between these characteristics and approaches to data reduction that could be applied to reduce an unwieldy number of variables.

Factor Analysis. Our work has been informed by lexical processing studies that explore multiple word-level factors to explain speed and accuracy in lexical decision and naming tasks. In this literature, researchers are often concerned with reducing the number of dimensions or variables that are used to predict performance. Clark and Paivio (2004) analyzed 32 word characteristics (e.g. length, frequency) for 925 nouns and found nine latent factors that explained 84% of the variance across variables. Brysbaert, Mandera, McCormick and Keuleers (2018) replicated this approach with even more discrete variables and found eight latent factors that explained 73% of the variance of the same set of words.

We opted to omit subjective, behavioral measures (e.g. reaction time, accuracy) in our data reduction analysis and instead focus on 19 discrete, objective variables (e.g. length, frequency; see Appendix 1). We conducted an exploratory factor analysis (EFA) with an oblique rotation in R to account for multi-collinearity between factors. An EFA was used so that we could explore the replicability of previous analyses without coercing variables to belong to any particular factor, and because our sample of words was strictly academic instead of words in general. We were able to extract five factors for our academic words, which explained 60% of the variance. The factor structure is presented in Figure 1, with only loadings and correlations more extreme than ±0.3 displayed. Table 1 then explains the working factor names and a general explanation of the factor. The five-factor model fit the data reasonably well (RMSR = 0.02, RMSEA = 0.087, TLI = 0.891, CFI = 0.941). Appendix 2 gives a breakdown of factor loadings for all variables.

Table 1. Factor names and explanations for word characteristics.

Factor	Variance Explained	Example Words	Numeric Meaning *Larger Number* =
(ML1) Orthographic Complexity	22%	proportion (3.8) consent (0) aid (-2.5)	more complex word
(ML3) Proximity	12%	role (12.1) random (0) dimension (-0.5)	more neighbors
(ML4) Frequency	11%	found (5.8) accompany (0) mature (-2.3)	more frequent
(ML2) Semantic Dispersion	10%	found (2.8) concept (0) schedule (-2.8)	more contexts
(ML5) Senses	5%	draft (2.5) create (0) dimension (-2.9)	more related meanings

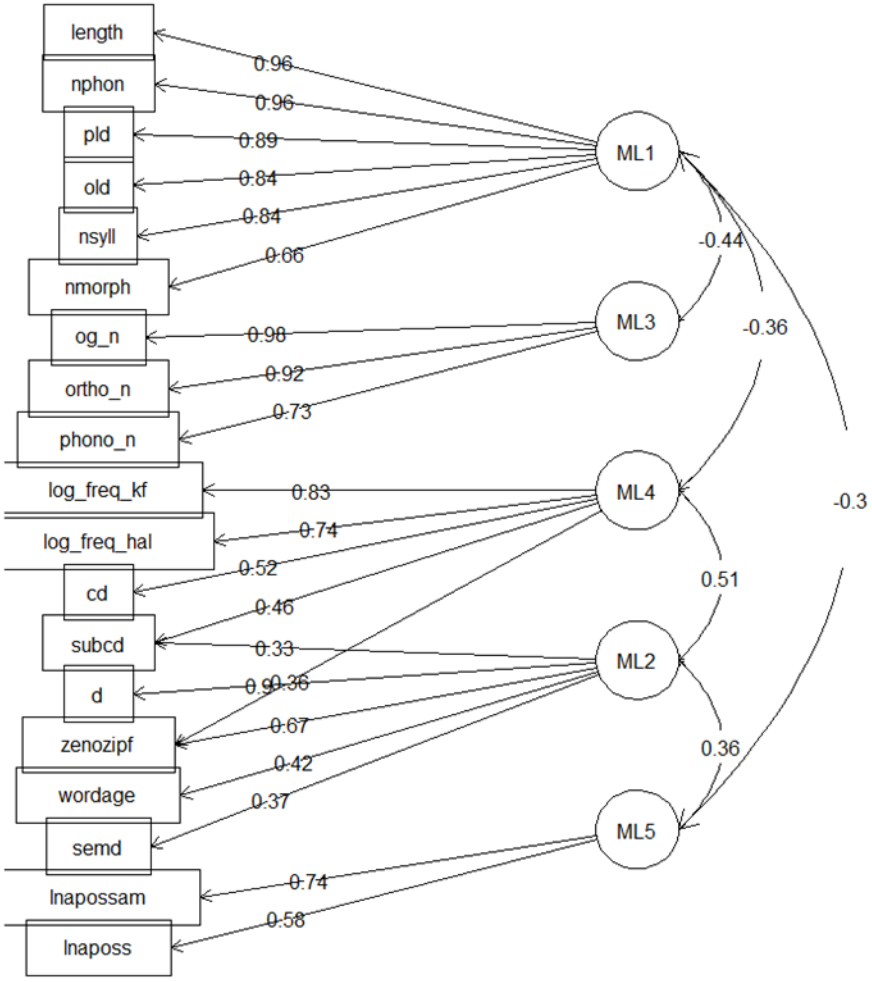


Figure 1. Factor structure of word characteristics

Differential Item Functioning, Distractor Functioning, and eIRT Analyses

We have conducted DIF analyses using three approaches (logistic regression, IRT, and generalized mixed linear model via GLIMMIX) as proposed in the grant application. In the Year 1 progress report we reported on results from the logistic regression analyses for uniform and non-uniform DIF, and also stated that we had begun conducting DIF analyses using the IRT approach. At the time of that writing, we described the IRT DIF analyses as "preferable" to the logistic regression analyses because IRT provides greater sensitivity and allows for analysis of items on both the 1011 and 1112 forms simultaneously using a common items equating approach to link the two forms due to the presence of 19 overlapping items. We did not present IRT results in the Year 1 report, because they were not yet complete. We have since completed those analyses, along with DIF analyses using the generalized mixed linear model approach using SAS 9.4 PROC GLIMMIX. The latter approach is the basis for eIRT, and thus a major step in the DIF analysis was to contrast the IRT estimates of DIF to the GLIMMIX estimates of DIF without any explanatory item characteristics in the models. We report on those analyses here, and the

challenge that ensued due to differences in the IRT and GLIMMIX estimates for selected items, namely the items that were unique to the 2011-2012 test form of Word Generation.

Table 2 provides item level descriptive information for each Word Generation Test item for both English Only and Limited English Proficient students. The extensive table includes the item number, the group, the target word being tested, the string of response options separated by "-" in their order of appearance on the test form (A-B-C-D) with the correct option in CAPS, which we designate the Key Word. The table also includes the number and percent of students in each group who selected each response option. In addition, the table includes two columns of information, Form_Use and ITEM_FORM, that indicate which forms the item appeared on, and the form and item number for the reported information. The column Form_Use encodes in the first letter whether the item is a Word Generation (W) item or Gates McGinitie (G) item. The next three elements are binary codes (0=No; 1=Yes) indicating whether the items appeared on the 2009-2010, 2010-2011, and 2011-2012 test forms. Thus, W111 is a Word Generation item that appeared on all three test forms, whereas an item with Form_Use = W101 is a Word Generation item that appeared in 2009-2010 and 2011-2012. Because the dataset only includes student data from years 2010-2011 and 2011-2012, such an item would only have data from the 2011-2012 school year, as is also the case for items coded W001. The problem, of course, with the information in Table 2 is that pass rates differ across groups because of ability differences between the groups as well as differential item functioning (DIF), if it exists. However, the table is useful as a description of basic group performance on each item and the attractiveness of different distractors for each item. The ITEM_FORM column provides information on the test form and item number on that form, with 1011_ indicating an item on the 2010-2011 test form and 1112 indicating an item on the 2011-2012 test form. The number following the underscore indicates the item number of the item on that test form to facilitate locating the item on the test form. Across the two test forms there are 110 items, only 80 of which are unique items.

Table 2 – Descriptive Statistics on Item Performance for Each Item for EO and EL Students.

In Year 1, we reported on estimation of DIF using logistic regression models. Here we focus on comparison of the logistic regression, IRT, and GLIMMIX estimates. It is important to understand that there are two equivalent IRT formulations for the one-parameter logistic model, and that IRT estimation software and GLIMMIX parameterize the models slightly differently. Specifically, IRT estimates a model where the slope parameter is constrained equal across items and groups, and the variance of the latent ability in the reference group (here EO's) is fixed at 1.0. This allows IRT to estimate the variance of the latent ability in the target group (here, ELs), as well as a mean difference in ability, while also estimating differences in difficulty parameters for each item between the target and reference group. Thus, IRT is estimating a model where the probability of a correct response is a non-linear function of $a(\Theta_i - b_j)$ where a is the discrimination (i.e., slope) parameter, constrained equal across items and groups, b_j is the difficulty parameter for item j, such that higher values of b indicate more difficult items, and Θ_i is the ability of individual i.[1] Importantly, this model can be reparameterized without loss of generality, such that the probability of a correct response is $\Theta_i + c_j$, where Θ_i is as before and c_j defines an item intercept that reflects item easiness. In this formulation, the slope parameter is forced to 1.0 and the intercept is $-b_j/a$. Thus, items with small intercept values are very difficult items and items with large intercept values are very easy. GLIMMIX and other generalized mixed linear model formulations use the latter parameterization, but estimate a slope parameter that is allowed to differ from 1.0, but is constrained equal across items. This difference introduces a change in the variance in ability in the reference group between the two estimation approaches, and essentially rescales the ability distribution between IRT and GLIMMIX.

To appreciate these differences, we plot the estimates for different parameters for the reference and target group in Figure 2.

Figure 2a. Comparison of IRT and GLIMMIX Estimates for Reference Group (EOs)

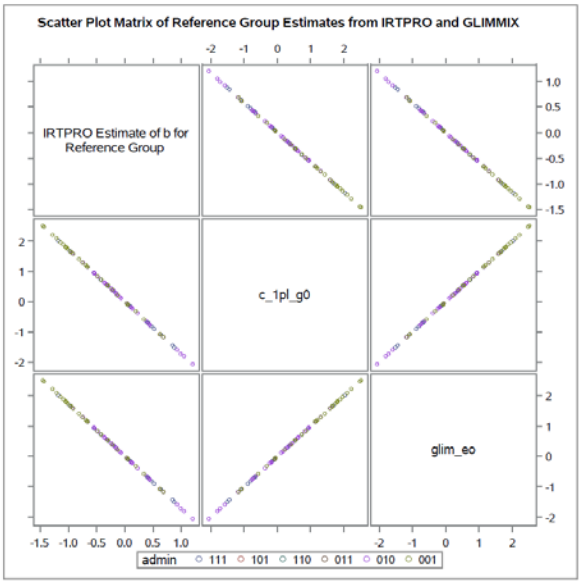
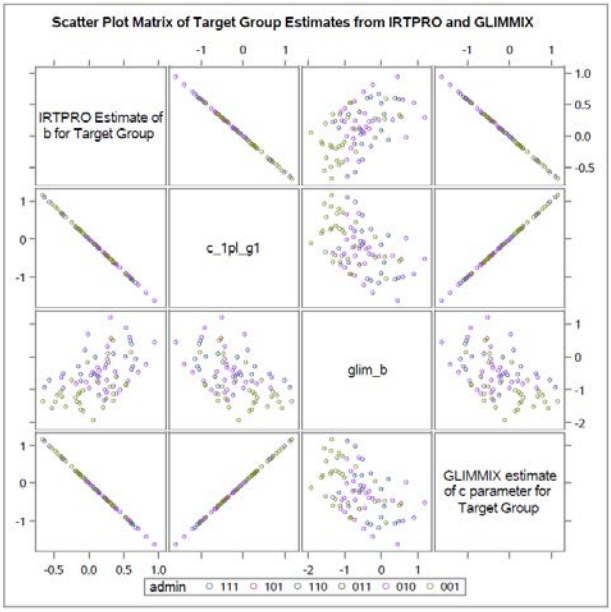


Figure 2b. Comparison of IRT and GLIMMIX Estimates for Target Group (ELs)



[1] Other IRT models are available that relax the assumption of constant slope across items, and that allow for guessing, but we have focused on application of the one-parameter model depicted here because of its desirable measurement properties. Discussion of these alternatives is beyond the scope of this report.

The scatterplot matrix of parameter estimates in Figure 2a compares the IRT estimate of b_j to the IRT estimate of c_j and the GLIMMIX estimate of the intercept parameter for the reference group. Clearly, there is perfect correspondence across the three estimates, but the conversion to intercept parameterization of the IRT model brings it into more direct correspondence with the GLIMMIX parameterization. When we examine the same scatterplot matrix for the target group in Figure 2b, things are similar. Here again, *b* and *c* represent the two different IRT formulations for the target group, and the expected, direct negative correspondence between estimates is apparent. The estimate glim_b is the coefficient for ELs in the GLIMMIX model, but this parameter is neither the *b* nor *c* parameter for the EL group, but rather represents the difference in *c* between the target group (ELs) and the reference group (EOs). When this model parameter is converted into the *c* parameter for the target group, we see direct correspondence with the IRT estimate of *c* and inverse relationship with the IRT estimate of *b*.

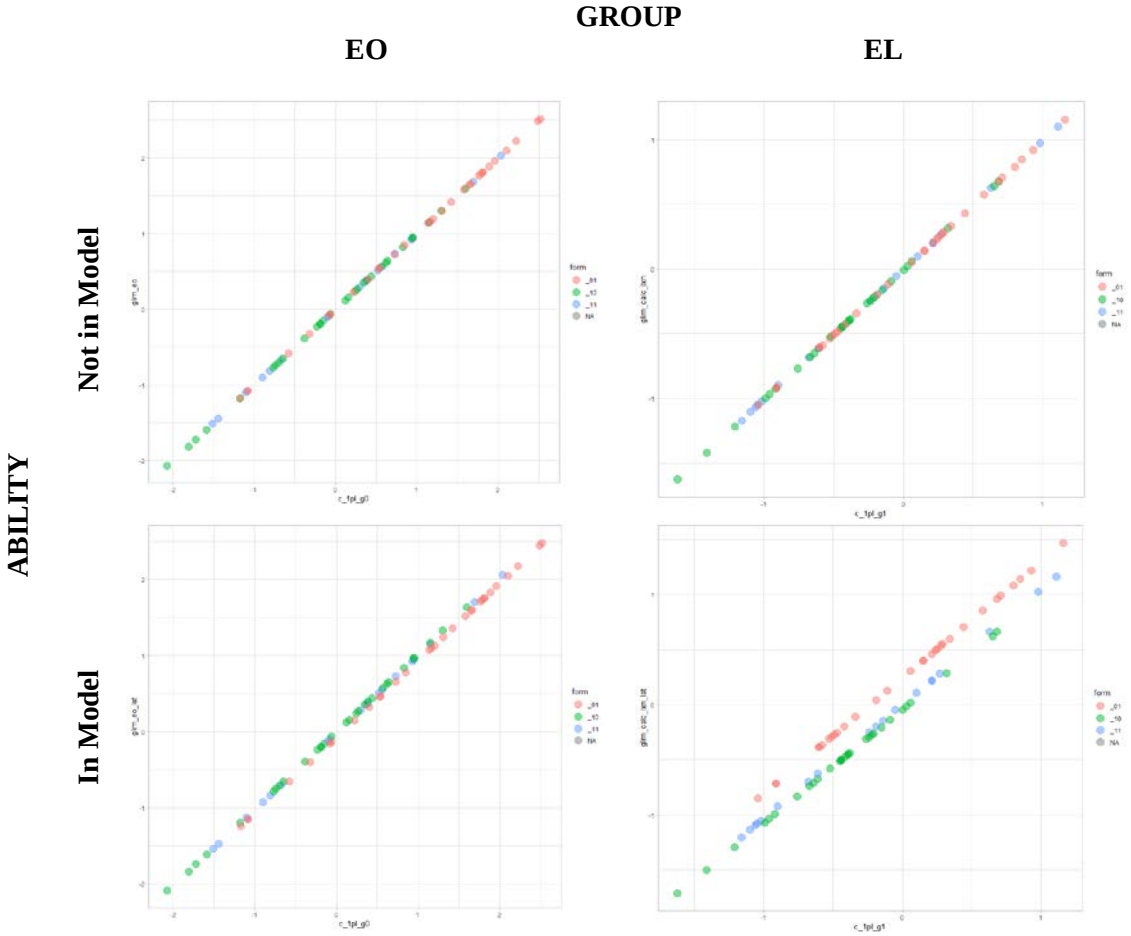
In the IRT model, the mean ability of the Reference group is set a 0 and the variance of the latent ability is set to 1.0, while the mean and variance of the target group are estimated in comparison to these values. In this instance, the mean for the ELs was estimated at -.36 and the variance at 0.28. These estimates imply that the standard deviation in the EL group is sqrt(.28) = .529 and the mean difference is -.36/.529 = .68 standard deviation units of the Target group. The GLIMMIX model constrains the mean of the Reference group at 0, and estimated the variance at 2.7055, while the mean and variance of the Target group (ELs) were estimated at -1.12 and .95, respectively. To place the GLIMMIX estimates on the IRT scale, we must divide through by the variance of the Reference group in the GLIMMIX model. Thus, the variance in the Target group from the GLIMMIX model on the IRT scale is .95/2.705 = .35, and the standard deviation is sqrt(.35) = .592, and the mean difference in ability is -.414. Expressed in terms of the standard deviation of the Target group, the difference of .414 amounts to a mean difference of .699 standard deviation units of the Target group, roughly the same as the .68 standard deviation unit difference observed from the IRT model.

The comparability of estimates across the two approaches appears very promising, but quickly becomes problematic when ability is introduced as a fixed effect in the model, which is necessary for the study of non-uniform DIF (i.e., the possibility that the difference between EOs and ELs is not uniform across the ability dimension, but rather is moderated by ability). When ability is explicitly coded into the model, differences emerge between the two modeling approaches.

These differences are apparent in Figure 3, which is divided into four panels in a 2 X 2 grid. In Figure 3, the relationship between *c* parameters for GLIMMIX (vertical axis) and IRT (horizontal axis) for EOs (Left Column) and ELs (Right Column) when ability is not in the model (Top Row), and when ability is explicitly included in the model (Bottom Row). When we examine the four panels, we see that the modeling approaches are equivalent for both groups when ability is not explicitly included in the model as a fixed effect, but only as a random effect of subjects (i.e., a latent mean, and variance of the latent ability). This model allows the mean ability to differ across groups by including group membership as a dichotomous variable, but ability is not included in the model as a measured predictor of performance. When ability is included as a fixed effect, we see a small difference in estimates for the Reference group (EOs), and a substantial difference across methods for the Target group (ELs). Most importantly, this difference between methods is not haphazard, but quite systematic, and specifically is linked to the test form.

We considered many possible causes for this difference, including the approach to ability estimation, the approach to scaling and/or centering the measure of ability in the model, the choice of anchor items for equating the scale across EL and EO students, etc. However, none of these elements was found to be responsible. Whether we estimated ability using a factor score from an analysis of both Gates and Word Generation vocabulary items, a measure from the Gates only, or a measure based only on the Word Generation test, the difference remained. It was not until we examined the effect as a function of test forms that the source of the problem became clear. Figure 3 color codes information about the test form into each panel. Specifically, the data points coded blue represent items that appear on both the 2010-2011 and the 2011-2012 test form, whereas the data points coded green represent items that appear only on the 2010-2011 test form. The data points coded red represent items that appeared only on the 2011-2012 test form. These are the items that are affected, but only for the EL group. Because these data points are shifted above the 45 degree reference line, they indicate that estimates of *c* are larger for ELs for items on the 2011-2012 test form when estimated using the GLIMMIX model as compared to the IRT model. Put another way, the IRT estimate makes these items appear more difficult for ELs and the GLIMMIX estimate makes them appear easier.

Figure 3 – Comparison of IRT (Horizontal Axis) and GLIMMIX (Vertical Axis) Estimates for *c* for EOs (Left Column) and ELs (Right Column) with (Bottom Row) and without (Top Row)



Note: This figure graphs IRT estimates of c (item easiness) on the horizontal axis against GLIMMIX estimates of c on the vertical axis. Plots in the left-hand column show estimates for EOs and plots in the right-hand column show estimates for ELs. Plots in the top-row are from the GLIMMIX model where ability is only a random effect in the model (i.e., a random student intercept), whereas plots in the bottom-row are from the GLIMMIX model where a latent measure of ability is also included as a fixed covariate effect in the model.

One might think that the problem is with the GLIMMIX estimate, but that, in fact, is not the case. This shift in *c* for items on the 2011-2012 test form for ELs occurs because the IRT model does not adequately adjust for the difference in ability between the two cohorts of ELs who are contributing data to the analysis. Because we restricted the analysis to the first time that students took the Word Generation assessment, the 2011-2012 cohort is comprised predominantly of Grade 6 students, and newly arriving students in Grades 7 and 8. These Grade 6 ELs and newly arriving Grade 7 and 8 EL students tend to be lower in ability than the overall cohort of EL students who participated in 2010-2011, in part because they are younger (i.e., predominantly Grade 6 students), but also because the newly arriving 7th and 8th grade students are different from the students who were in the schools the preceding year.

This difference in ability is apparent in Table 3 below, which provides descriptive statistics by grade for EL and EO students in the 2010-2011 and 2011-2012 cohorts involved in

the analysis (i.e., first time test takers). It is most apparent by comparing the mean performance on F and GMPV_TE, the latent ability estimate derived from a factor model of all vocabulary items and the Gates Vocabulary Extended Scaled Score, respectively, because these scores are on the same scale for all students. For the EO students, the latent mean increases from -.037 to .189 to .315 from Grade 6 to Grade 7 to Grade 8 in the 2010-2011 cohort, but decreases from .156 to -.12 to -.51 in the 2011-2012 cohort. The Gates scaled score mean shows a similar pattern increasing from 521 to 530 to 540 across grades in the 2010-2011 cohort and declining from 521 to 517 to 509 in the 2011-2012 cohort. For the EL students, the differences across grades indicate an overall decline from 2010-2011 to 2011-2012, but with improved performance as a function of grade in 2010-2011 and worsening performance across grades in 2011-2012.

Table 3 – Descriptive Statistics for EO and EL Students by Grade and Cohort

group	School Year	School Grade	N Obs	Variable	Label	N	Mean	Std Dev	Minimum	Maximum
0	1011	6	982	F		982	-0.0371324	1.0035040	-2.4850000	2.3600000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	910	25.8252747	10.5846277	0	49.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	965	26.4735751	9.6699998	4.0000000	45.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	965	521.2538860	41.3948049	405.0000000	653.0000000
		7	967	F		967	0.1885119	1.0259535	-2.4850000	2.6690000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	909	28.5599560	11.0397750	1.0000000	50.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	934	25.7494647	10.0523482	0	45.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	934	530.3511777	40.5388255	372.0000000	661.0000000
		8	921	F		921	0.3152052	1.0977634	-2.4850000	2.6690000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	874	30.0961098	12.0178336	0	50.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	877	28.0239453	9.9744249	1.0000000	45.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	877	540.2029647	41.2028044	383.0000000	661.0000000
	1112	6	614	F		614	0.1563550	0.9360985	-2.7730000	2.4400000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	596	31.6963087	10.5079427	0	50.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	602	26.6262458	9.4860354	0	45.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	602	521.5382060	41.4149816	354.0000000	653.0000000
		7	70	F		70	-0.1208286	1.2351241	-2.7730000	1.4970000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	63	30.5873016	12.1318495	1.0000000	46.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	69	22.4492754	9.9580535	3.0000000	44.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	69	517.0289855	40.0796894	406.0000000	632.0000000
		8	46	F		46	-0.5075870	1.5174472	-2.7730000	1.8620000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	36	29.5277778	12.7177853	0	48.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	44	20.4772727	11.7225456	1.0000000	45.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	44	509.0681818	55.0705246	383.0000000	661.0000000
1	1011	6	532	F		532	-0.6665883	0.6274656	-2.4850000	0.9590000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	500	15.1760000	5.7182533	1.0000000	38.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	515	14.7359223	5.4697764	4.0000000	33.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	515	474.1398058	23.3357299	405.0000000	543.0000000
		7	518	F		518	-0.5579846	0.6907777	-2.4850000	1.9360000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	486	16.8930041	6.8184764	0	47.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	490	13.8224490	6.1772644	0	36.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	490	483.3000000	29.9098475	372.0000000	563.0000000
		8	469	F		469	-0.4393987	0.7016694	-2.4850000	1.7770000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	450	18.1800000	7.9079363	0	46.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	441	15.4421769	6.6395467	1.0000000	45.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	441	491.1950113	31.1193169	383.0000000	661.0000000
	1112	6	243	F		243	-0.7569630	0.6955174	-2.7730000	1.6630000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	235	18.4808511	7.9862766	0	47.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	242	14.3181818	5.8183071	3.0000000	36.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	242	471.6198347	26.2519728	394.0000000	557.0000000
		7	52	F		52	-0.7819423	0.8268148	-2.7730000	1.3540000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	49	18.6326531	9.5605393	5.0000000	45.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	50	12.9200000	8.1637326	1.0000000	44.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	50	478.9000000	39.9179004	383.0000000	632.0000000
		8	37	F		37	-1.1058108	0.9670182	-2.7730000	0.6460000
				WGV_CTR	WORD GENERATION VOCABULARY: Corrected Raw Score	30	17.9666667	8.3148531	5.0000000	38.0000000
				GMPVCTR	GATES MACGINITIE: VOCABULARY: Corrected Total	37	11.4054054	6.1393967	0	30.0000000
				GMPV_TE	GATES MACGINITIE: VOCABULARY: ESS - Extended Scale Score	37	469.5135135	34.7799604	372.0000000	545.0000000

Note: Group 0 = EO Students; Group 1 = EL Students; F = Latent Ability Factor Score from Word Generation and Gates Vocabulary Items; WGV_CTR = Word Generation Number Correct; GMPVCTR = Gates Vocabulary Raw Score; GMPV_TE = Gates Vocabulary Extended Scaled Score

Specifically, the latent mean increases from -.666 to -.558 to -.439 from Grade 6 to 7 to 8 in 2010-2011, but is worse in 2011-2012 and declines from -.75 to -.78 to -1.1 from Grade 6 to 7 to 8. Similarly, the Gates scale score increases from 474.1 to 483.3 to 491.2, but is lower in each grade and decreases from 471.6 to 478.9 to 469.5 in 2011-2012.

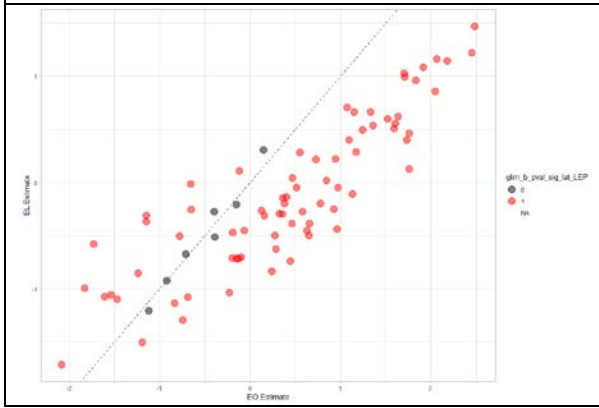
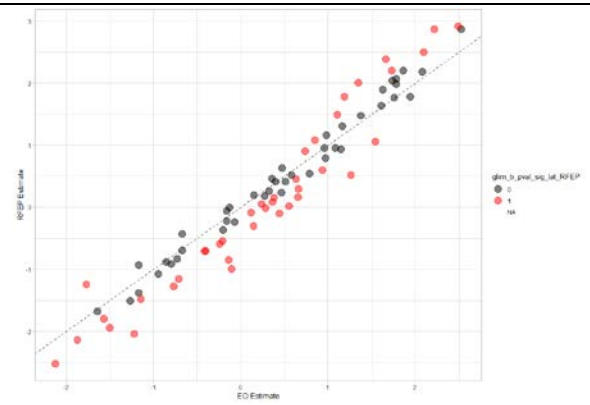
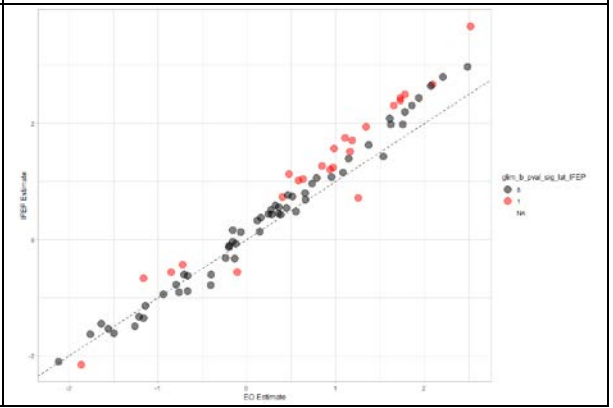
By including an explicit estimate of ability in the model, the GLIMMIX model adjusts for these ability differences between the groups of students providing data on the different sets of items for EOs and ELs. The small difference in ability in the two EO cohorts (.152 vs .088) produces a small difference across methods for the items that are unique to the 2011-2012 test form, that is further diminished by the fact that the 2011-2012 cohort has a mean value closer to 0 (i.e., the latent mean of the ability distribution for EOs). In contrast, the difference in ability is relatively large between the two EL cohorts (-.559 vs -.800) as measured by the latent factor means, especially in light of the smaller standard deviation among the ELs compared to the EOs, and the mean of the 2011-2012 cohort is further from 0 than the mean of the 2010-2011 cohort. These factors combine to exacerbate the magnitude of DIF estimated by the IRT approach for the items unique to the 2011-2012 test form because the ability difference between ELs and EOs taking those items are not adequately controlled in the IRT model.

We spent a considerable amount of time getting to the bottom of this problem, and ultimately proved the case by simulating a dataset where there was no DIF between groups. In our simulation, the number of items is similar to the real data, and there are the same number of items that are unique to each form and shared between the two forms. Most importantly, although there is no DIF, there were ability differences between groups and cohorts that paralleled our ability group differences. When we analyze this simulated data using IRT and GLIMMIX with and without ability estimates in the model, we corroborate the results in Figure 3. Thus, on balance we infer that the GLIMMIX estimates with ability in the model are the best estimates of actual DIF, whereas the IRT estimates and GLIMMIX estimates without ability in the model are potentially biased due to cohort ability differences.

Using estimates of DIF from the GLIMMIX models with a factor score based on all Gates and WG items in the model as an estimate of ability, we see that most items on the WG test show evidence of DIF favoring the EO's (n = 60), whereas a handful of items (n = 13) seem to favor the ELs. Only seven items did not show evidence of DIF based on the standard error of the Group*Item interaction in the GLIMMIX model. In contrast, most items did not show evidence of DIF when comparing EOs and language minority students who were identified as initially fluent English proficient (IFEP). Specifically, 24 of 80 items showed some evidence of DIF, only three of which favored EO's, and 21 of which favored the IFEPs. When comparing recently reclassified ELs (RFEP) to EOs, we find 11 items showing DIF that favors the RFEPs and 27 items showing DIF that favors the EOs. The remaining 42 items did not show statistically significant DIF. Additionally, it is clear that item easiness/difficulty differentially relates to DIF across these subgroups. Specifically, it is the most difficult items (those with low "easiness" estimates) that tend to favor the ELs, whereas easier items tend to favor the EOs. It is NOT the case that the more difficult items tend to be cognates. It is possible that these more difficult items have attractive distractors that tend to be more attractive to EOs, thereby lowering the probability of getting the item correct by chance. At this point, we have to consider this explanation as very speculative, as we have not yet examined possible explanations. Still, in contrast to the items showing DIF for EL-EOs, items throughout the range of difficulty performed comparably for EOs and IFEPS, and items favoring the IFEPS tended to be in the upper half of easiness. Three items with below average easiness favored the IFEPS. In contrast,

the three items favoring EOs were spread across the easiness dimension from moderately easy, to about average, to very difficult (viz., the second most difficult items for EOs showed DIF favoring EOs. In strong contrast to the items showing DIF between EO and EL students, the items showing DIF that favored RFEPs tended to be above average in easiness. Only one item with below average easiness that showed DIF favored the RFEPs, whereas 11 items favoring RFEPs were above average in easiness. Items favoring the EOs existed across the range of item easiness from moderately easy (i.e., upper quartile in easiness) to very difficult (i.e., 6 of the 11 most difficult items for EOs showed DIF favoring EOs, while 1 showed DIF favoring RFEPs, and 4 did not show DIF). To assist in visualizing these relationships, we provide Figure 4 which graphs item easiness for EOs against item easiness for each of the target groups. In Figure 4, red points represent items showing DIF and gray points represent items not showing DIF. Points above the diagonal line signal items that favor the Target Group (i.e., items that are easier for the Target group), whereas points below the diagonal line favor EOs.

Figure 4. – Item Easiness Estimates and DIF for Target Groups (ELs, RFEPs, and IFEPs) and Reference Group (EOs).

DIF for ELs vs EOs	DIF for RFEPs vs EOs	DIF for IFEPs vs EOs
		

Note: Red points indicate items showing DIF; gray points indicate items not showing DIF. Red data points above the line indicate items that are "easier" for the Target Group (i.e., ELs, RFEPs, or IFEPs), and red data points below the diagonal line indicate items that are "easier" for the Reference Group (i.e., EOs). The horizontal axis represents item easiness for EOs and the vertical axis represents item easiness for the Target group (i.e., ELs, RFEPs, or IFEPs from left to right, respectively).

Explanatory Item Response Models for Item Easiness and DIF

We have begun preliminary analyses of explanatory item response models by ordinary least squares regression and correlation of item characteristics with item parameter estimates from GLIMMIX models. This approach gives us access to tools of regression for descriptive statistics regarding possible effects, as well as model and variable selection tools available in regression that are not available in full blown, multivariate, cross-classified random effects models of eIRT. These models are also faster to estimate and have been advocated as an approach to identification of potential covariates and predictors in multi-level modeling since the advent of the HLM software. Presently, we have examined models based on the five factors of item characteristics described above in the first section of the Year 2 report, but have also looked at specific characteristics loading on these factors, and have begun doing best variable subset regression models for modeling of item easiness and direct modeling of DIF. In Table 4, we provide correlations among item characteristics and item easiness estimates for different groups.

Table 4 – Correlations of Item Characteristics with Item Easiness for Target and Reference Groups

Item Characteristic	Correlations with DIF Estimates for Group			Correlations with Easiness Estimates for Group			
	LEP	IFEP	RFEP	EO	LEP	IFEP	RFEP
d	**0.36355** **0.0009**	**0.32233** **0.0035**	**0.30939** **0.0052**	**0.3377** **0.0022**	-0.19627 0.081	0.1702 0.1312	0.07971 0.4822
wordage	0.10652 0.347	0.17899 0.1121	0.14468 0.2004	0.18115 0.1078	-0.20849 0.0635	0.11846 0.2953	-0.03791 0.7385
semd	**0.25587** **0.0228**	**0.25103** **0.0256**	**0.23273** **0.039**	**0.26144** **0.0199**	-0.18177 0.1089	0.13843 0.2237	0.03601 0.7527
subcd	**0.49671** **<.0001**	**0.53479** **<.0001**	**0.53067** **<.0001**	**0.55651** **<.0001**	**-0.44284** **<.0001**	**0.29641** **0.0076**	0.21049 0.0609
zenozipf	**0.40399** **0.0002**	**0.43543** **<.0001**	**0.43557** **<.0001**	**0.41638** **0.0001**	**-0.2936** **0.0082**	**0.37762** **0.0006**	**0.30401** **0.0061**
og_n	-0.06575 0.5623	-0.0211 0.8526	-0.02445 0.8296	-0.00555 0.961	-0.06647 0.558	-0.07182 0.5267	-0.07414 0.5134
ortho_n	-0.00266 0.9813	0.02751 0.8086	0.01473 0.8968	0.04695 0.6792	-0.08934 0.4307	-0.05196 0.6471	-0.09568 0.3985
phono_n	0.12819 0.2571	0.10942 0.3339	0.12687 0.2621	0.12443 0.2715	-0.07903 0.4859	0.02169 0.8486	0.07859 0.4883
log_freq_kf	0.19003 0.0956	0.2189 0.0542	0.21186 0.0626	0.19903 0.0806	-0.14306 0.2115	**0.22762** **0.045**	0.15885 0.1648
log_freq_hal	**0.2798** **0.012**	**0.31101** **0.005**	**0.3102** **0.0051**	**0.31754** **0.0041**	**-0.25691** **0.0214**	0.195 0.083	0.14823 0.1894
cd	**0.278** **0.0125**	**0.3083** **0.0054**	**0.31142** **0.0049**	**0.27122** **0.015**	-0.17393 0.1228	**0.35477** **0.0012**	**0.30323** **0.0063**
lnapossam	**0.32592** **0.0032**	**0.41035** **0.0002**	**0.39094** **0.0003**	**0.43451** **<.0001**	**-0.41793** **0.0001**	0.19995 0.0754	0.07566 0.5047
lnaposs	**0.31382** **0.0046**	**0.32478** **0.0033**	**0.31747** **0.0041**	**0.35797** **0.0011**	**-0.29148** **0.0087**	0.10591 0.3498	0.04467 0.694
first_appear	-0.15111 0.1809	-0.1003 0.3761	-0.14049 0.2139	-0.10998 0.3315	0.02577 0.8205	-0.03469 0.76	-0.177 0.1163
bg_mean	**-0.25903** **0.0203**	-0.19954 0.076	**-0.23333** **0.0373**	**-0.22323** **0.0465**	0.10792 0.3407	-0.05265 0.6428	-0.16217 0.1507
old20	-0.06429 0.5685	-0.08591 0.4457	-0.10221 0.3639	-0.12209 0.2776	0.14956 0.1827	0.07471 0.5074	0.00811 0.9427

Note: Elements in bold are statistically significant at p<.05.

We have fit a number of models explaining item easiness and DIF, and we wish to highlight one of those models. Specifically, of great interest are the possible joint effects of word frequency, contextual diversity, and multiple word senses (polysemy). Results of a direct model of item easiness predicted from these predictors and their interactions with group membership (i.e., DIF) are presented in Table 5.

Table 5 – Explanatory Model Showing Combined Individual Predictors from Factors of Item Characteristics

		ITEM EASINESS		
	Variable	***b***	***SE***	***p-val***
Main effects	zenozipf	-0.027	0.170	0.876
	subcd	0.517	0.174	0.004
	lnapossam	0.229	0.111	0.043
Intercepts	IFEP	0.531	0.126	<.001
	LEP	-0.164	0.069	0.020
	RFEP	0.251	0.125	0.048
	EO	0.332	0.102	0.002
Interactions	zenozipf*IFEP	0.119	0.055	**0.032**
	zenozipf*LEP	0.067	0.099	0.502
	zenozipf*RFEP	0.127	0.064	0.051
	subcd*IFEP	-0.006	0.056	0.919
	subcd*LEP	-0.240	0.101	**0.021**
	subcd*RFEP	-0.010	0.066	0.876
	lnapossam*IFEP	0.030	0.036	0.408
	lnapossam*LEP	-0.150	0.065	**0.023**
	lnapossam*RFEP	-0.002	0.042	0.971

Note: zenozipf – Zipfian transformation of word frequency from Zeno et al.; subcd – sub contextual diversity; lnapossam – number of word senses. IFEP – dichotomous indicator equal to 1 for IFEP students; LEP - dichotomous indicator equal to 1 for EL students; RFEP - dichotomous indicator equal to 1 for RFEP students

Table 5 shows that, on average, words are easiest for IFEP students, then EOs, and then RFEP students. Words are most difficult for EL students. Contextual diversity and the number of senses both impact word easiness, such that word easiness increases as contextual diversity and the number of senses increases. Importantly, these effects are not uniform across groups. Specifically, both contextual diversity and the number of senses have a less beneficial effect for EL students. In addition, word frequency tends to increase word easiness for IFEP and RFEP students, but not for ELs or EOs. We provide graphs of the bivariate relationships between predicted values from the multiple regression analysis and the individual predictors in Figure 5 below. The figure does not depict the multiple regression (i.e., unique contribution) slopes, but the bivariate relation between the text feature and expected item easiness given the predictors.

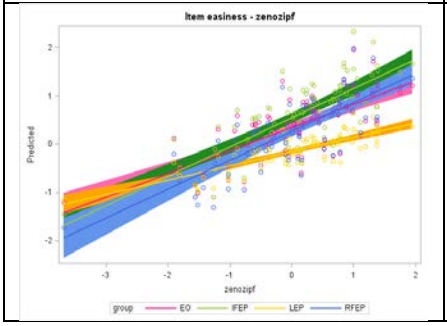
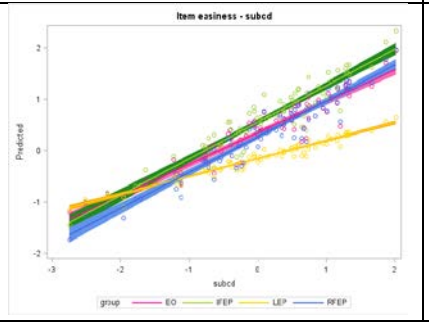
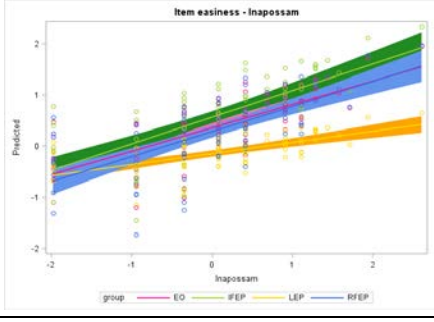
Our preliminary investigations into item characteristics that affect item difficulty and DIF are interesting in light of research on word meanings and word learning. In research on lexical

decision tasks (LDT), words with multiple senses (i.e., polysemous words) tend to be identified faster, which has been referred to as the "ambiguity advantage" (Azuma & Van Orden, 1997; Balota, Ferraro, & Connor, 1991; Borowsky & Masson, 1996; Eddington & Tokowicz, 2015; Hino & Lupker, 1996; Millis & Button, 1989; Pexman & Lupker, 1999; Rubenstein, Garfield, & Millikan, 1970; Woollams, 2005; Yap et al., 2011). Researchers have hypothesized that words with multiple related senses might be represented more accurately as having a single meaning that is flexible with respect to interpretation in different contexts. That is, the meanings of these words may fall within a single large semantic space that *facilitates* network performance and thus lexical retrieval (Rodd, Gaskell, & Marslen-Wilson, 2004).

On the other hand, there is evidence that words with multiple unrelated meanings (i.e., homonyms) are processed less efficiently in LDTs. This finding suggests that there is an efficiency penalty for adjudicating the possible meanings that a particular word form represents, resulting in less efficient lexical retrieval. The parallel distributed processing (PDP) model of ambiguity processing predicts both the penalty for multiple meanings and the facilitation effect for word forms with multiple senses. One of the challenges for this model is operationalizing the difference between meanings and senses. Rodd, Gaskell, and Marslen-Wilson point out that "etymological, semantic, and syntactic" criteria have been used to define this distinction and that "there may not always be a clear distinction between these types of ambiguity" (p. 246). This distinction is even more opaque in natural reading environments, in which document-level features influence meaning selection. For example, the word *product* only has one meaning, but the sense in math texts is distinct from the one found in history texts even though they are logically and etymologically related.

Despite the interest in meanings and senses in the lexical processing literature, there are relatively few studies examining how the number of meanings may influence student performance on vocabulary assessments, or how the number of meanings a word has relates to item ease for students across levels of language proficiency. Our findings in this area could highlight an important issue in both the teaching, learning, and assessment of vocabulary, especially given that "high leverage" general academic words tend to have more meanings than other words.

Figure 5 – Graphs of Interaction of Word Characteristics with Item Easiness Across three Target Groups

Word Frequency	Contextual Diversity	Number of Senses
		

Note: Each panel depicts the bivariate slope relating item easiness to a specific text feature (Word Frequency, Contextual Diversity, and Number of Senses) for each of four groups. The red line shows the relationship for the reference group (EOs), the green line = IFEP, the orange line = ELs, and the blue line = RFEP. The color band around each line provides a 95% confidence interval on the regression slope. The depicted lines are bivariate slopes, not slopes from the multiple regression analysis. Multiple regression slope effects were as follows: for Number of Senses and Contextual Diversity the slope of the line for ELs is flatter than the slope of the line for EOs,

whereas the slopes of the lines for IFEP and RFEP are not different from the slope of the line for EOs. For Word Frequency, the slope for ELs is not different than the slope for EOs, whereas the slope for RFEPs and IFEPs is different than the slope for EOs.

Opportunities for Training

Nothing to Report

How were the results communicated to Communities of Interest?

During the second reporting period we presented as part of a symposium at the Annual Meeting of the American Educational Research Association in Toronto in April, and are presenting as part of a symposium at the Annual Meeting of the Society for the Scientific Study of Reading in July of 2019. We are working on manuscripts to report on our analyses to date and are developing a website that we expect to be operational by the end of grant year 2 (August 31, 2019). We are also submitting as part of two symposia to present at AERA in San Francisco next April.

II. PRODUCTS

No papers have been published to date. We anticipate one or more papers to be submitted on DIF, our factor analysis of item characteristics, and the e-IRT by the end of this calendar year. We had hoped to prepare multiple papers during the second reporting period, but complications with the DIF analyses slowed our progress, but will ultimately lead to a valuable publication. Other products further developed in year 2 are the build out of the various databases of item characteristics. These separate, linked databases are now largely complete in terms of the features we had originally planned to integrate along with numerous additional features. We have identified several "item-level" features that we are considering adding that involve characteristics of multiple elements of an item, rather than characteristics of the target word or key word. For example, the average word frequency of the target, key, and distractors, or the variance in number senses across target, key, and distractors. Addition of these features is a low priority at this time, but we are considering ways to improve the prediction of item easiness and DIF that capture elements of the constellation of the item, and not just the target and key word characteristics. We anticipate making elements of the database available to the public, along with the SHINY App, which allows exploration of the link between item features and item easiness, as well as DIF on the website when it is launched later this summer.

III. Participants & Other Collaborating Organizations

KEY PERSONNEL

Name:	Dr. David J. Francis
Effort	1 month
Role	PI/PD
Location	TIMES, University of Houston, Houston, TX, United States

Description	Directs the overall program of research, provides guidance on statistical modeling and analyses, coding, study design and analysis. Overall responsibility for budget, managing and approving expenditures, and maintaining study timelines and communication with IES. (Dr. Francis effort is paid in the summer, which is not included in this initial reporting period and thus his funding is not reflected in the SF 424 Expenditure Report).
Name:	Dr. Paulina A. Kulesz
Effort	2.6 months
Role	Co-Investigator
Location	TIMES, University of Houston, Houston, TX, United States
Description	Dr. Kulesz is responsible for carrying out the analyses associated with DIF, DDF, and e-IRT.
Name:	Dr. Autumn McIlraith
Effort	12 months
Role	Post-Doctoral Fellow
Location	TIMES, University of Houston, Houston, TX, United States
Description	Dr. McIlraith is a post-doctoral fellow working on the project and is based at TIMES at the University of Houston. She works with Drs. Francis, Kulesz, and Lawrence on the coding of item features, especially those characteristics of a linguistic or orthographic nature, such as LSA, LD, and OLD20, and is asisting in the design, organization, and compilation of the database, as well as in the design and construction of tables and graphical displays of results in order to communicate findings as clearly as possible. Dr. McIlraith will also assist Drs. Francis and Kulesz in the DIF, DDF, and e-IRT analyses.
Name:	Dr. Joshua Lawrence
Effort	1 day per week
Role	Co-Investigator
Location	Lawrence Consulting, MA, United States
Description	Dr. Lawrence has directed the collection and coding of information on words, word meanings, and items. Dr. Lawrence has taken the lead on the identification of key features to code for words and items as well as directing the design of the database and has directed the extraction of information from the relevant databases that are publicly available. He is also the lead contact with the state database that contains information on student characteristics and is a member of the team that developed the Word Generation Academic Vocabulary Assessment. He and Dr. Francis collaborated on the original Word Generation Efficacy Trial that was funded by IES and that is providing the student data for the current project in Phase I. Dr. Lawrence's effort is contracted through Lawrence Consulting. He is also a faculty member at the University of Oslo and spends the majority of the calendar year in residence in Oslo, Norway.

He does spend some extended time each calendar year in the US. The team communicates weekly with Dr. Lawrence via email and through a weekly conference call hosted on Webex/Zoom. He regularly attends these meetings, regardless of his location. Dr. Lawrence also supervises Ms. Rebecca Knopf, a graduate student at the University of Oslo, who has worked on the project (viz., contributing to the development of the item features database and conducting the factor analysis of word characteristics). Ms. Knopf is not supported financially on the grant.

OTHER PERSONNEL

Name:	Martin Walczak
Effort	25% effort for three months
Role	Graduate Student Research Assistant
Location	TIMES, University of Houston, Houston, TX, United States
Description	Mr. Walczak worked with Drs. Kulesz and McIlraith on the coding of words based on the number of definitions they have. He has no access to student data.

OTHER SUPPORT FOR KEY PERSONNEL

As proposed in the report for Year 1, rather than hire a second post-doc, we increased the effort of Dr. Lawrence to 1 day per week. There have been no other changes to the other support for the PD/PI or other Key Personnel since the Just in Time information was submitted.

PARTICIPATING ORGANIZATIONS

University of Houston, Houston, TX

IV. IMPACT

The project has had no measureable impact to date as we are still in data compilation and analysis mode in Phase I of the work. We are not aware of any citations of the work that was presented earlier this year at AERA.

V. CHANGES/PROBLEMS

The only significant challenge that we encountered was the lack of correspondence between IRT and GLIMMIX estimates for item easiness and DIF, which led to significant additional effort in data analysis. This problem stemmed from our decision to restrict the 2011-2012 cohort to students without prior experience with the Word Generation test. Because we expected that the

IRT results were correct, we spent a disproportionate amount of time trying to "correct" the GLIMMIX estimates, only to determine that the GLIMMIX estimates were correct and it was the IRT estimates that were biased for the items that were unique to the 2011-2012 test form, but only for the Target Group and not for the Reference Group (EOs). We have been meeting successfully virtually every week as a team via Webex/Zoom and made good progress on our objectives for Year 2, although we did not get to the point of piloting items. We expect to be able to pilot items in the fall and still collect data in grant year 3, as originally planned. We feel that the project remains on track as currently being implemented.

VI. SPECIAL REPORTING REQUIREMENTS

No special reporting requirements were listed in the Notice of Grant Award or the Performance Agreement.

VII. BUDGETARY INFORMATION

See attached SF 424 for the Expenditure Report from the Reporting Period

BUDGET JUSTIFICATION FOR GRANT YEAR 2 ACTIVITIES

Key Personnel:
David J. Francis serves as the Principal Investigator on the project. He will direct and oversee the project, statistical analyses related to the project aims, and statistical reporting. In particular, he will supervise the work of Dr. Kulesz who will implement the explanatory item response models, differential item functioning, and differential distractor functioning analyses, and Dr. Autumn McIlraith, post-doctoral Fellow working on the project. Funds are requested to cover one month (8%) of his summer salary and fringe in each year of the project.

Paulina A. Kulesz serves as co-Investigator on the project. She will work with Dr. Francis to oversee any data management tasks that are necessary to prepare the data for analysis. Dr. Kulesz will be primarily responsible for implementing the explanatory item response models, differential item functioning, and differential distractor functioning analyses. She will also assist Drs. Francis and Lawrence in scientific reporting. Funds are requested to cover 4.8 months (40%) of her salary and fringe in each year of the project.

Dr. Joshua Lawrence, Lawrence Consulting and the University of Oslo, serves as co-Investigator. Dr. Lawrence's role on the project is described more fully under Consultant Services, reflecting how he will be paid on the project. However, his role is that of co-Investigator, reflecting his central role in the proposal. When the project was originally submitted, Dr. Lawrence was at the University of California at Irvine. Since then he has moved to the University of Oslo, thus making it necessary to engage Dr. Lawrence through a consulting services contract while serving as a co-Investigator. His effort in year 2 of the project will be 2 calendar months.

Dr. Autumn McIlraith, Post-doctoral Fellow, was hired in year 1 of the project as a post-doctoral Fellow at TIMES, University of Houston. She is employed full-time on the project in year 2 to assist Dr. Kulesz with statistical analyses in the second year of the project.

Other Personnel:

Laudemer Vigilia, Data manager, will be responsible for organizing, controlling, and aggregating data. He is supported for two months on the project.

Martin Walzak, is a research assistant who has worked with Drs. Kulesz and McIlraith on the coding of words based on the number of definitions they have. He has no access to student data.

Project Manager 1 (to be determined) at 8% effort (1 calendar month) in year two. This person will be in charge of examiner training and the coordination of data collection at the schools. This person will maintain a data collection management system that will organize and register the staff, materials and sites involved in data collection. In Year 2, only pilot testing of new items is proposed for new data collection, so it is not anticipated that we will hire and train examiners. Rather, data collection will be handled by the project manager and student volunteers. This position was not hired due to the fact that we have not yet begun pilot testing.

Fringe Benefits:
Fringe benefits are based on actual amounts and calculated using a Fringe Benefits Calculator supplied by UH Office of Contracts and Grants.

Travel:
Funds were requested to support travel by Drs. Francis and/or Kulesz to the required IES meetings in year 2, and for key personnel to present findings from year 1 activities at professional conferences. In terms of travel to the IES meetings, the estimated cost per person per trip will be $1,367 in year one and will include airfare, lodging, ground transportation, meals and incidentals. In terms of travel to a conference (including escalation in each year), the cost per person per trip was estimated at $2,465 and included airfare, lodging, ground transportation, meals and incidentals, and conference registration fees. Funds were also requested to support travel to Houston for the 4 consultants at $3,832. Costs will include airfare, lodging, ground transportation, meals and incidentals. We have not convened a meeting of the consultants in Houston.

Other Direct Costs:

Consultants Services: Funds are requested to cover a consultant rate of $1000 per day per consultant in each year of the project. The consultants are Drs. Mikyung Wolf, Catherine Snow, Young-Suk Kim, and Paul DeBoeck. We are budgeting to cover the consultant rate for 4 consultants who will consult on the content, substantive nature of obtained findings, and statistical analyses. Depending on the needs, we are proposing to meet with the consultants 1 to 2 days in the coming year.

Dr. Joshua Lawrence, Co-Investigator: Funds are requested for Dr. Josh Lawrence. Dr. Lawrence is listed as a consultant because of his relocation from UC Irvine to the University of Oslo during the period of time in between submission of the proposal and its being awarded. Dr. Lawrence will commit two months of effort to the project in year 3. Dr. Lawrence will serve as co-Investigator on the overall project, will work with Dr. Francis in directing all aspects of the work conducted, including development of the item coding for existing items, integration of new language information regarding existing items, and development of item models DIF analyses, and DIF distractor analyses, planning and conducting the pilot data collection, statistical analyses of pilot data, and the development of new items, and leading the preparation of presentations, reports, and publications about project findings.

Indirect Costs:
Indirect costs are calculated on a modified total direct costs basis using the DHHS-approved rate of 50.5%.

References

Azuma, T., & Van Orden, G. C. (1997). Why SAFE Is Better Than FAST: The Relatedness of a Word's Meanings Affects Lexical Decision Times. *Journal of Memory and Language*, *36*(4), 484–504. https://doi.org/10.1006/jmla.1997.2502

Balota, D. A., Ferraro, F. R., & Connor, L. T. (1991). On the early influence of meaning in word recognition: A review of the literature. In P. Schwanenflugel (Ed.), *The Psychology of Word Meaning* (pp. 187–222). Hillsdale, NJ: Lawrence Erlbaum Associates.

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(1), 63–85. https://doi.org/10.1037/0278-7393.22.1.63

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2018). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-1077-9.

Clark, J. M., & Paivio, A. (2004). Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers: A Journal of the Psychonomic Society, Inc*, *36*(3), 371–383. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/15641426

Eddington, C. M., & Tokowicz, N. (2015). How meaning similarity influences ambiguous word processing: the current state of the literature. *Psychonomic Bulletin & Review*, *22*(1), 13–37. https://doi.org/10.3758/s13423-014-0665-7

Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology. Human Perception and Performance*, *22*(6), 1331. https://doi.org/10.1037/0096-1523.22.6.1331

Millis, M. L., & Button, S. B. (1989). The effect of polysemy on lexical decision time: now you see it, now you don't. *Memory & Cognition*, *17*(2), 141–147. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/2927312

Pexman, P. M., & Lupker, S. J. (1999). Ambiguity and visual word recognition: can feedback explain both homophone and polysemy effects? *Canadian Journal of Experimental Psychology = Revue Canadienne de Psychologie Experimentale*, *53*(4), 323–334. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10646204

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modeling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*(1), 89–104. https://doi.org/10.1016/j.cogsci.2003.08.002

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487–494. https://doi.org/10.1016/s0022-5371(70)80091-3

Woollams, A. M. (2005). Imageability and ambiguity effects in speeded naming: convergence and divergence. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(5), 878–890. https://doi.org/10.1037/0278-7393.31.5.878

Yap, M. J., Tan, S. E., Pexman, P. M., & Hargreaves, I. S. (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review, 18*(4), 742–750. https://doi.org/10.3758/s13423-011-0092-y

Appendix 1

Factors, Variables, and Related Information for Item/Word Characteristics

Factor	Variable	Dataset	Meaning	Numeric Meaning (Bigger Number = …)
Orthographic Complexity	length	Wordnet	Number of letters in word	longer word
	nphon	ELP	Number of phonemes (distinct sounds)	longer word (more sounds)
	pld	ELP	Phono**log**ic Levenshtein distance 20 – distance (in number of steps) from the word to the 20 closest Levenshtein neighbors	more complex word (more steps to closest neighbors)
	old	ELP	Orthographic Levenshtein distance 20 – distance (in number of steps) from the word to the 20 closest Levenshtein neighbors	more complex word (more steps to closest neighbors)
	nsyll	ELP	Number of syllables	longer word (more syllables)
	nmorph	ELP	Number of morphemes (distinct meaningful units)	more complex word
Proximity	og_n	ELP	Number of phono**graph**ic neighbors a word has	closer to other words (common)
	ortho_n	ELP	Number of orthographic neighbors a word has	closer to other words (common)
	phono_n	ELP	Number of phono**log**ical neighbors a word has	closer to other words (common)
Frequency	log_freq_kf	ELP (Brown)	Word frequency in the Brown corpus (log transf)	more frequent word
	log_freq_hal	ELP (HAL)	Word frequency (log transf) in HAL corpus, which is from usenet?	more frequent word
	cd	ContDiv (Adelman)	Contextual diversity; number of documents in a coprus that contain that word	more documents
	subcd	Subtlex	Contextual diversity; number of documents in a coprus that contain that word (standardized)	more documents
	zenozipf	Zeno	Word frequency (zipfian transf)	more frequent
Semantic Diversity	d	Zeno	Number of content areas where word appears, regardless of frequency (log transf)	more content areas
	wordage	Google ngram	Number of years word has existed (as of 2000) (transformation of firstyear)	word has existed long (is older)
	semd	SemD (Hoffman)	Diversity of nearest semantic neighbors (log transf) using LSA as a proxy for how semantically diverse the word itself is	more diverse contexts
Senses	lnapossam	Wordnet	Number of senses and meanings across all parts of speech	more senses and meanings
	lnaposs	Wordsmyth	Number of senses across all parts of speech	more senses

Note: ELP – English Lexicon Project; Google ngram – frequency database of letter strings; Wordnet, ContDiv, SubtLex, Zeno, Wordsmyth, SemD are defined in Year 1 Progress Report (see Appendix 3). These are provided immediately below as well for ease of access.

Word Net Meanings and Semantic Precision

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Mill will er, K. J. (1990). Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, *3*(4), 235–244. https://doi.org/10.1093/ijl/3.4.235

The WordNet project resulted in the creation of a relational database connecting more than 200,000 English words in a relational database. One of the unique features of this database is that it is keyed on mea ning (synsets) rather than word forms. This data coding process forced coders to attend carefully to multiple meanings for words, and researchers using these data have found that there are more individual senses and meanings of words listed on average than you would find in a typical dictionary (such as Wordsmyth.com). We merged our initial set of words with a large data set extracted with a Python program resulting in a data set of 6,221 meanings for our words. Average number of meaning varied across part of speech; *M* adj = 2.8, *M* noun = 3.8, *M* adverb = 1.8, *M* verb = 4.23).

Wordnet data are organized in hypernym changes. For example, animal > dog> poodle. We successfully merged the depth of each word meaning to each original meaning. On average nouns were more precise in this metric (M = 6.8) than verbs (M = 2.2). Even controlling for frequency, we think specificity may be a marker for utility that varies by language proficiency.

ELP - English Lexicon Project

Lexical Access

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B.,& Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17958156

ContDiv

Contextual Diversity

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times: Supplemental File. *Psychological Science*, *17*(9), 814. Retrieved from http://www.adelmanlab.org/cd/cdpstabs.pdf

subtlex

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

zeno

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, N.Y.: Touchstone Applied Science Associates.

Ngram

https://walshbr.com/textanalysiscoursebook/book/issues/google-ngram/

SemD

Semantic Diversity

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

wordsmyth

https://www.wordsmyth.net/

Appendix 2

Factor Loadings, Communalities, and Unique Variances

Variable	Factor Loadings					Variances	
	ML1 ***Ortho***	**ML3** ***Prox***	**ML4** ***Freq***	**ML2** ***SemD***	**ML5** ***Senses***	**h2**	**u2**
cd	0.04	0.10	0.52	0.23	0.00	0.455	0.545
lntotalmeanings	0.00	0.23	-0.01	-0.01	0.10	0.067	0.933
lnaposs	-0.02	0.01	0.11	-0.19	0.59	0.334	0.666
d	-0.04	-0.01	-0.05	0.90	-0.01	0.764	0.236
zenozipf	0.00	0.03	0.36	0.68	0.04	0.873	0.127
semd	-0.01	-0.01	0.02	0.37	0.02	0.146	0.854
length	0.96	0.00	-0.04	0.01	0.08	0.911	0.089
log_freq_hal	-0.15	0.00	0.75	0.00	0.05	0.681	0.319
ortho_n	-0.03	0.92	-0.04	0.02	-0.01	0.854	0.146
phono_n	-0.12	0.73	0.05	-0.03	-0.01	0.634	0.366
og_n	0.06	0.98	0.02	-0.01	0.00	0.913	0.087
old	0.84	-0.06	0.05	-0.02	-0.11	0.801	0.199
pld	0.89	-0.02	0.06	-0.05	-0.09	0.845	0.155
nphon	0.96	0.00	-0.01	-0.02	0.07	0.890	0.110
nsyll	0.84	-0.02	0.01	-0.01	-0.08	0.757	0.243
nmorph	0.66	0.05	-0.24	0.11	0.07	0.528	0.472
log_freq_kf	0.00	0.01	0.82	0.02	0.01	0.695	0.305
subcd	0.03	0.04	0.45	0.33	0.14	0.541	0.459
lnapossam	-0.05	0.01	-0.01	0.10	0.72	0.612	0.388
freqband	0.06	-0.08	0.28	0.10	-0.07	0.103	0.897
wordage	0.03	-0.03	-0.03	0.43	0.22	0.275	0.725

Note: h2 - Communality; u2 - Uniqueness

Appendix 3

Detailed Report on Year 1 Accomplishments

Details of Item Coding

During the first reporting period, we made progress on coding item level characteristics. We used items from the Word Generation Academic Vocabulary Assessment, as well as items from several forms of the Gates Vocabulary test. We then expanded out to include different inflected forms and different possible meanings of the target and response words from this broad set of items. We merged the expanded word set with a number of different sources of lexical information (see section A below). These lexical features will serve as predictors in later analyses of item functioning, as well as provide us with a set of words from which to develop future test items. Several additional features were coded only for the original target and response words from the Word Generation test (section B below). These features required more time and effort to code and the Word Generation test words were treated as the priority. These additional features may be coded for the expanded word set as well in the future, if we determine that they are informative. Table 2 below contains a visual representation of the main item characteristics that have been coded up to this point.

A. Expanded Word Set

The purpose of this milestone is to bring together relevant item level characteristics for all target words and distractors used in any of the assessments that provided data for this project. We took a rather broad perspective on what might constitute a relevant characteristic, consulting research in linguistics, reading, child development, and psychology. The norms for processing lexical data were not always consistent across these sources, and as a result we could not not always achieve perfect merge rates with our target words. We briefly document each source and describe the success of the merge below. In all cases, we create flag variables documenting the success/failure of the merge at the link-level so that we are certain of the reason for any missing data, and this information is maintained in the database.

Our final database includes data on 6,221 meanings of 1,023 word forms. There are a total of 269 variables in our current database, and that does not include some important orthographic and phonological controls that we will also use in our analysis. This information will be incorporated at a later time. The complete list of variables including means, min, max and label is in section II. PRODUCTS.

<u>Initial set of words</u>

We began with 1,023 words culled from the Word Generation Academic Vocabulary Assessment, and Gates Reading and Vocabulary tests (forms 6 and 7/9). This count only includes singleton target words; we have developed a plan for coding and evaluating multiword expressions which exist in our target assessments. We will describe the merge success and analysis of these data in our next progress report.

<u>Word Net Meanings and Semantic Precision</u>

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Mill will er, K. J. (1990). Introduction to

WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, *3*(4), 235–244. https://doi.org/10.1093/ijl/3.4.235

The WordNet project resulted in the creation of a relational database connecting more than 200,000 English words in a relational database. One of the unique features of this database is that it is keyed on meaning (synsets) rather than word forms. This data coding process forced coders to attend carefully to multiple meanings for words, and researchers using these data have found that there are more individual senses and meanings of words listed on average than you would find in a typical dictionary (such as Wordsmyth.com). We merged our initial set of words with a large data set extracted with a Python program resulting in a data set of 6,221 meanings for our words. Average number of meaning varied across part of speech; *M* adj = 2.8, *M* noun = 3.8, *M* adverb = 1.8, *M* verb = 4.23).

Wordnet data are organized in hypernym changes. For example, animal > dog> poodle. We successfully merged the depth of each word meaning to each original meaning. On average nouns were more precise in this metric (M = 6.8) than verbs (M = 2.2). Even controlling for frequency, we think specificity may be a marker for utility that varies by language proficiency.

Frequency

Davies, M. (2015). The Wikipedia Corpus: 4.6 million articles, 1.9 billion words. *Adapted from Wikipedia. Accessed February*, *15*.

- We needed to collapse raw frequency across the detailed POS variable used by Davies to complete this merge, but we ended up getting a merge rate of 95.26%.

Age of Acquisition

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

- Kuperman used the uninflected forms to generate the AOA estimates, so we had to merge on lemma value from Davies.
- Final merge rate was 90.79%, but we might want to treat estimate of inflected forms carefully.

Semantic Diversity

Hoffman, P., Lambon Ralph, M. A., & Rogers, T. T. (2013). Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730. https://doi.org/10.3758/s13428-012-0278-x

- Merge rate was 97.99%

Contextual Diversity

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times: Supplemental File. *Psychological Science*, *17*(9), 814. Retrieved from http://www.adelmanlab.org/cd/cdpstabs.pdf

- Merge rate was 99.76%

Frequency & Diversity in School Texts

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency*

guide. Brewster, N.Y.: Touchstone Applied Science Associates.

- Merge rate was 99.98%

Lexical Access

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B.,& Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17958156

- Merge rate was 99.97%

Frequency

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

- Merge rate was 99.04%

Valence, arousal and domincance

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

- Merge problems we do the fact that that using data set only had 13K words so many of our words had no norms.
- Merge rate was 67.27%

Concretenss

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

- Merge problems due the fact that that the master dataset is composed of many extremely abstract word that were not selected by the Brysbaert team for this study
- Merge rate was 61.40%

Imaginability

Bird, H., Howard, D., & Franklin, S. (2003). Verbs and nouns: the importance of being imageable. *Journal of Neurolinguistics*, *16*(2–3), 113–149. https://doi.org/10.1016/S0911-6044(02)00016-7

- Merge problems due to the fact that the master dataset is composed of unimaginable words that were not selected by the Bird team for this study
- Merge rate was only 20.19%

B. Word Generation Word Set

The Word Generation test has three forms, administered in years 2009-2010, 2010-2011, and 2011-2012. Each form contains 50 items, but some items are common across forms. There are 110 unique items across the three forms. Eighty-one of these unique items appear on forms 2010-2011 and 2011-2012, the years for which we have student-level data in the current study.

Item features were coded across the entire set of 110 items, to enable the inclusion of the full set of items in the future measure development phase of the study. Each item consists of a carrier sentence, in which one word is underlined, and four response options. The student is instructed to choose which of the response options is closest in meaning to the underlined word. We subsequently refer to the underlined word as the target, the correct response as the key, and the incorrect responses as the distractors. Several orthographic and semantic features were coded for the set of 550 words from the 110 items, which will be tested as explanatory variables in later analyses (see Table 2 for a summary).

Levenshtein distance.

We investigated orthographic similarity between words using Levenshtein distance (Yarkoni, Balota, & Yap, 2008). The Levenshtein distance (LD) between two words is the minimum number of operations that are needed to turn one word into another (operations consisting of substitution, insertion, and deletion). For example, ROSE to NOSE would have a LD of 1, and CHANCE to STRAND would have a LD of 5 (3 substitutions, 1 insertion, 1 deletion). Because it goes beyond the number of immediate neighbors that can be created with a single operation, this metric provides a more continuous metric of similarity and can be used to examine the density of orthographic neighborhoods in a more flexible manner than more constrained approaches.

LDs were calculated for each item, between the target word and each of the response options, as well as between the key and each of the distractors. This information was also summarized per item, yielding measures of the average LD between target and responses, average LD between key and distractors, and average LD between target and distractors. Most of the targets and responses consisted of single words. For the multi-word targets and responses, we removed the spaces between words and treated them as a single string. This resulted in several strings which were outliers in terms of length, and LDs relative to these strings will be excluded if they are determined to be unduly influential in the analyses. We calculated all LDs using the R package *vwr* (Keuleers, 2013).

When the test was created, an attempt was made to select distractor words such that one distractor had a high degree of orthographic similarity to the target word. We found that for some of the items, this orthographically similar distractor could be identified on the basis of its low LD from the target. However, for some items, the key was in fact more similar to the target than any of the distractors. We are interested to investigate the impact of this key-target distance on item difficulty and item functioning in future analyses. In addition, the similarity of the distractors to the target will be of interest when examining distractor functioning in future analyses.

OLD20

The mean LD from a word to its 20 closest neighbors is a measure of the density of that word's orthographic neighborhood. This metric was created by Yarkoni and colleagues (2013). As the distances between a word and its neighbors increase, the OLD20 metric will increase, such that higher values indicate a sparser neighborhood or more isolated word. OLD20 values were calculated for the target word and response options for each item, using the *vwr* package. This variable is a useful supplement to the LD, since it examines each word independently rather than the similarity between pairs of words.

Citations for Levenshtein distance and OLD20:

Keuleers, E. (2013). vwr: Useful functions for visual word recognition research. R package version 0.3.0. https://CRAN.R-project.org/package=vwr

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review, 15*(5), 971–979.

Latent Semantic Analysis: One-to-many, term to term comparison

Latent semantic analysis (LSA) is a computational model which maps the meanings of texts and words within a constructed semantic space. Meanings of texts are computed as a function of the meanings of all the words contained in the text, and word meanings are computed based on all the occurrences of the word across texts. Given a large corpus of texts, LSA creates a semantic space, or map, and represents each word and text as a vector within that space. Similarities between the meanings of words or texts can then be calculated based on the similarity of their vectors (Landauer, 2007).

We calculated similarities between the target word and each response option, as well as between the key and the distractors, using a web interface created by researchers at CU Boulder (lsa.colorado.edu), using a one-to-many comparison. We calculated similarities based on two different semantic spaces available on the website: *General reading up to 9th grade*, and *General reading up to 1st year college*. Thus for each item, there were 14 comparisons: 4 comparisons of the target item to each of the response options and 3 comparisons of the key to each of the distractors, within each semantic space. Most of the targets and response options consist of a single word. The multi-word targets and responses were not an issue for LSA, which is flexible and can accept words, phrases, and even entire texts.

Use of the web interface requires that each set of words be pasted into a window separately. One coder calculated all of the similarities, and an independent coder checked 5% of the items. There was 100% agreement between the first and second coders. Using the 9th grade semantic space, there were 7 items whose targets did not appear in the space and therefore target-response similarities could not be calculated. Four items had keys that were not present and therefore key-distractor similarities could not be calculated. Using the 1st year college semantic space, no items had targets that were not present. Two items had keys that were not present. A number of items had one or more response words missing, in one or the other semantic spaces. LSA similarities were therefore not calculable for all the word pairs, but the amount of missing information is relatively small.

Citation:

Landauer, T. K. (2007). LSA as a theory of meaning. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.). *Handbook of latent semantic analysis*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Table 2. Orthographic and Semantic features that were coded for each word within each item (single values), or for pairs of words (relative values). Sources are indicated within the green boxes. Some of these features were also coded for a larger set of words, including words from the Gates and various inflected forms of the original words.

	Orthographic		Lexical		Semantic			
	OLD20	**Levenshtein Distance**	**Frequency**	**Lexical Access**	**Semantic Precision**	**Age of Acquisition**	**Diversity**	**Latent Semantic Analysis**
SINGLE VALUES								
target key distractor 1 distractor 2 distractor 3 average of 5 words average of 4 responses average of 3 distractors	vwr package		Davies; SUBTLEX-UK; Zeno	The English Lexicon Project	WordNet	Kuperman	Semantic Diversity (Hoffman); Contextual Diversity (Adelman)	
RELATIVE VALUES								
target to key target to distractor 1 target to distractor 2 target to distractor 3 key to distractor 1 key to distractor 2 key to distractor 3 average of target to 4 responses average of target to 3 distractors average of key to 3 distractors		vwr package						LSA one-to-many (lsa.colorado.edu)

Examination of DIF

We made substantial progress in examining differential item functioning (DIF) of Word Generation (WG) items for ELL and English-Only (EO) students using the logistic regression approach. ELLs and EOs represented focal and reference groups, respectively. The logistic regression analyses were separately computed for each WG item within two WG test forms that were administered in the 2010-2011 school year (1011) and 2011-2012 school year (1112). The analyses for the two school years involve non-overlapping samples, and each sample consists only of those students taking the assessment for the first time (i.e., the individual student's first occasion of assessment). Consequently, the analysis of the 1112 test form is based on a smaller sample than the 1011 test form, because many students participating in 1112 also participated in 1011, but were excluded from the analysis of the 1112 form in the logistic regression analyses. The analyses were separately computed for each test form because (a) the two test forms included many non-overlapping items (31 items per test form), and (b) item numbering of overlapping items was not consistent across the two test forms. Analyses were also split into three sets depending on the type of total test score being used as a vocabulary knowledge ability estimate (the WG test score, Gates test score, or a latent score estimate of ability). The latent score was estimated using a 1PL confirmatory factor analytic model with 81 WG items (31 unique items from the WG 1011 test form, 31 unique items from the WG 1112 test form, and 19 overlapping items across the two test forms) and 90 Gates items (45 items from the Gates vocabulary subset for Grade 6, and 45 items from the Gates vocabulary subset for Grades 7-9). Conducting analyses of fifty items from two WG test forms, three ability scores, and three separate logistic regression models for each item (described below), we estimated a total of 900 logistic regression models. The three separate logistic regression models that were estimated for each item were, in fact, identical for each item, test form, and the ability estimate being used. Specifically, we estimated three different models:

Model 1 included the intercept and vocabulary ability.

$$Y = \beta_0 + \beta_1 VocabularyAbility \quad (1)$$

Model 2 included the intercept, vocabulary ability, and EL Group.

$$Y = \beta_0 + \beta_1 VocabularyAbility + \beta_2 ELGroup \quad (2)$$

Model 3 included the intercept, vocabulary ability, EL Group, and the interaction of vocabulary ability with EL Group.

$$Y = \beta_0 + \beta_1 VocabularyAbility + \beta_2 ELGroup + \beta_3 VocabularyAbility*ELGroup \quad (3)$$

To determine the presence of **uniform DIF**, we tested whether (a) the β_2 coefficient in Model 2 was significantly different from 0, and whether (b) that Model 2 provided a better fit relative to Model 1 by comparing a χ^2_{21} statistic

$$\chi^2_{21} = 2ln\,\frac{L(Model\ 2)}{L(Model\ 1)}$$

to the value from the χ^2 distribution with degrees of freedom equal to the difference in degrees of freedom between Models 1 and 2.

To determine the presence of **non-uniform DIF**, we tested whether (a) the β_3 coefficient estimated in Model 3 was significantly different from 0, and (b) that Model3 provided a better fit relative to Model 2 by comparing a χ^2_{32} statistic

$$\chi^2_{32} = 2ln\,\frac{L(Model\ 3)}{L(Model\ 2)}$$

to the value from the χ^2 distribution with degrees of freedom equal to the difference in degrees of freedom between the two models.

Figure 1 presents relations between different estimates of uniform DIF for each item on the 1011 test form under Model 2. The difference between the two estimates stems from the use of either the Gates as the measure of Vocabulary Ability or a latent score estimate of Vocabulary Ability. The two panels in the figure graph the exact same data points; the two panels differ in that the left-hand panel classifies items based on Model 2 using the Gates and the right-hand panel classifies items based on Model 2 using the latent ability estimate. Although the estimates for uniform DIF for a given item were highly correlated across the different approaches to estimating Model 2 (Note the overall positive relation in the scatter plot), the DIF classification for a given item varies between the left-hand panel and the right-hand depending on the choice of ability estimate. In Figure 1, the left hand side of the figure shows the classification of items when ability is estimated using the Gates score. The DIF classification based on the model with the Gates ability score suggested that some items favored EOs (blue dots in the lower left-hand corner), some items functioned the same for both groups (red dots), and some items favored ELs (blue dots in the upper right-hand corner). The DIF classification based on the model with the latent ability score was in line with investigators' expectations and suggested that most items showing DIF favored EOs (blue dots in the lower left-hand corner) while other items function the same for both groups (red dots). When the latent ability estimate is used, only one item favored ELs, the lone blue dot in the upper right corner of the scatter plot in the right-hand panel. This item measured knowledge of the word "recite". This item may have functioned differently due to its position on the WG assessment (the first item on the WG 1011 test form). While one might speculate about cognates and other word/item features that could account for the item favoring ELs, one must also keep in mind that most of the ELs in the present study are not Spanish speaking, given that the sample comes from northern California. Although the sample includes many Spanish-speaking students, they are not the majority language group among the ELs in this sample.

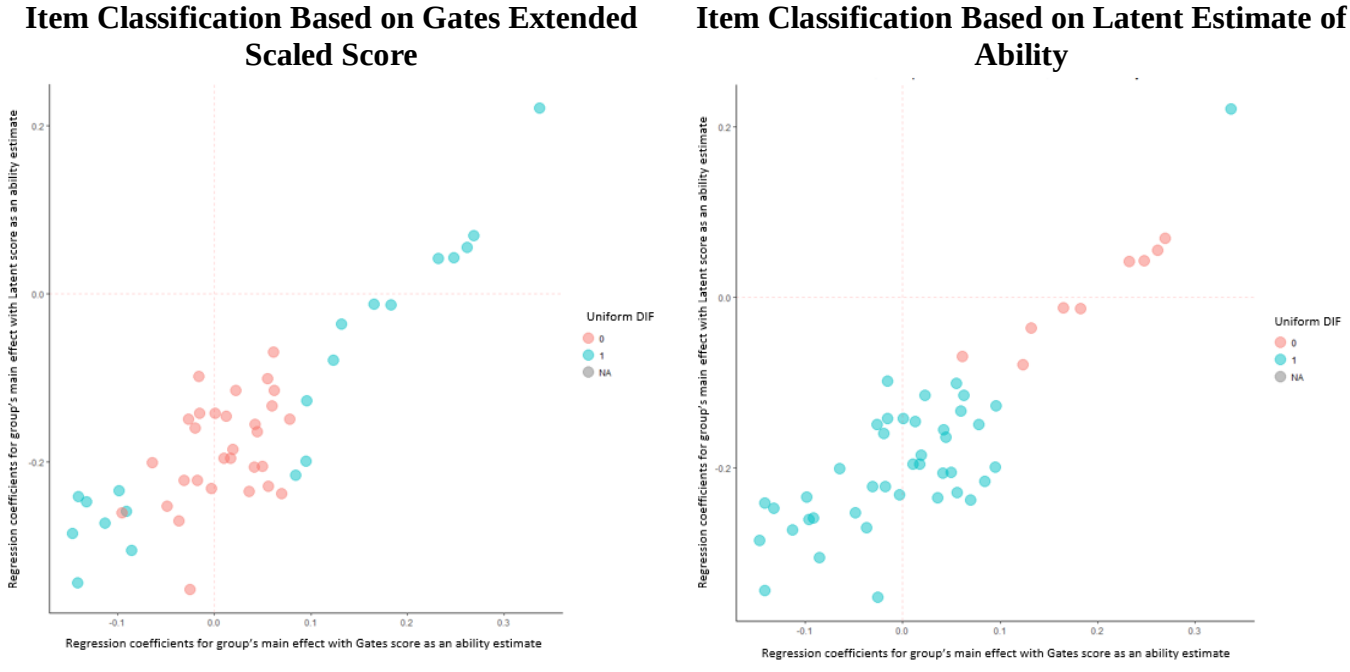


Figure 1. Each scatterplot demonstrates the relation between regression coefficients for main effects of group derived from Model 2 using either the Gates score (X axis) or the latent score (Y axis) as the measure of Vocabulary Ability in Model 2. ***The color-coding of data points in the panels indicates if the item was identified as showing uniform DIF. In the left-hand panel, color-coding is based on Model 2 estimated using the Gates Extended Scaled Score as an ability estimate. In the right-hand side, color-coding is based on Model 2 estimated using a latent score estimate of ability. In each panel, red dots represent items with no DIF. Blue dots represent items with uniform DIF.***

Figure 2 presents relations between two sets of estimates of non-uniform DIF, that is, it presents the relation between (a) the regression coefficients for the interaction effects (group by ability) for the fifty WG 1011 test items when estimated in Model 3 using the Gates score as a an ability estimate, and (b) the regression coefficients for the interaction effects for the fifty WG 1011 test items when estimated in Model 3 using the latent score as an ability estimate. Again, the two etimates of the regression coefficients were highly correlated. However, DIF classification for a given item varied based on the specific choice of ability estimate. The DIF classification based on the model with the Gates Extended Scaled Score as the ability estimate suggested that few items functioned the same for both groups (red dots in the lower left-hand corner), with the exception of one item that had a stronger relation with ability for EOs. This item measured knowledge of the word "apathy". The unusual functioning of this item may stem from the fact that a correct response alternative for this item was the word "unconcerned" which has a very low frequency of use. At the same time, the majority of items suggested stronger relations with ability for ELs (blue dots in the middle and upper right-hand corner). The DIF classification based on Model 3 using the latent ability score showed that ability was more strongly related to performance for EOs for some items (blue dots in the lower left-hand corner), that some items functioned the same for both groups (red dots), and that ability was more strongly related to performance for ELs for some items (blue dots in the upper right-hand corner).

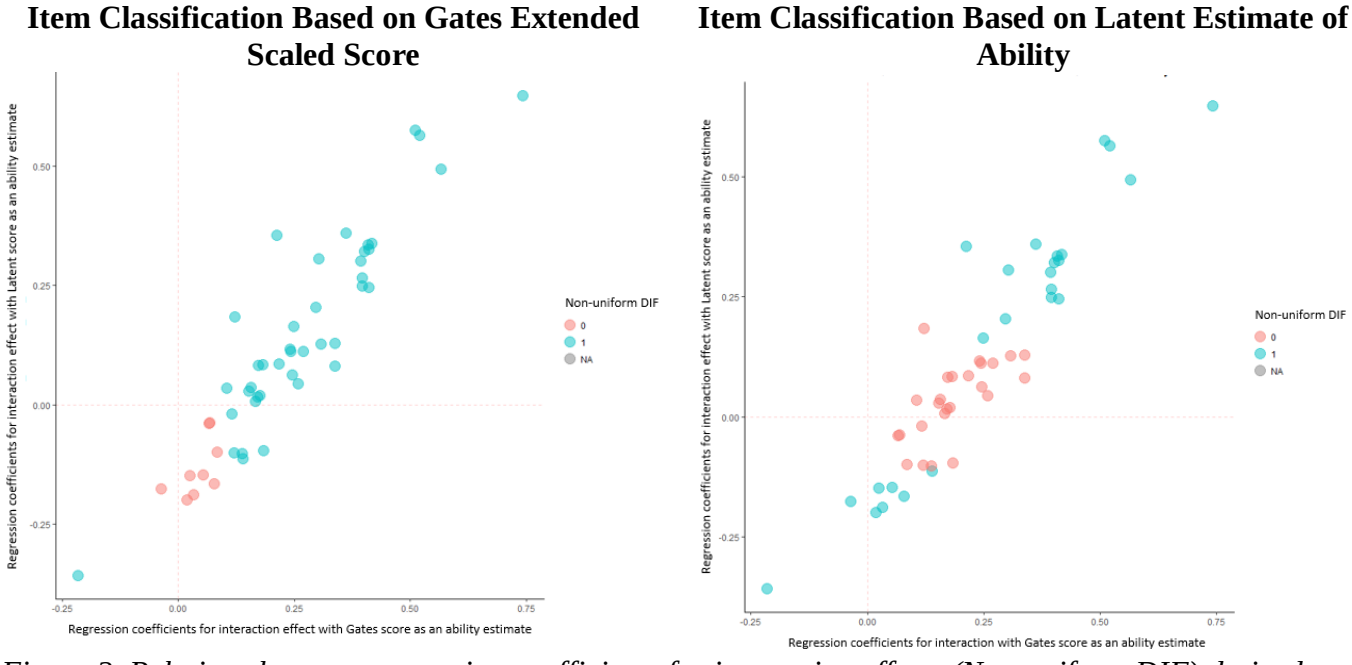


Figure 2. Relations between regression coefficients for interaction effects (Non-uniform DIF) derived from models using the Gates score (X axis) and latent score (Y axis) as the ability estimate in Model 3. . ***The color-coding of data points in the panels indicates if the item was identified as showing non-uniform DIF. In the left-hand panel, color-coding is based on Model 3 estimated using the Gates Extended Scaled Score as an ability estimate. In the right-hand panel, color-coding is based on Model 3 estimated using a latent score estimate of ability. In each panel, red dots represent items with no DIF. Blue dots represent items with non-uniform DIF.***

Patterns of relations between regression coefficients for main effects and relations between regression coefficients for interaction effects were generally similar for different ability scores within the WG 1011 and 1112 test forms. As presented in Table 3, the correlations between regression coefficients of main effects estimated using different ability scores within the WG 1011 and 1112 test forms ranged from .88 to .94. Similarly, the correlations between regression coefficients for interaction effects estimated using different ability scores within the WG 1011 and 1112 test forms ranged from .67 to .98.

Table 3. Correlations between regression coefficients for group and interaction effects for different ability estimates within the WG 1011 and 1112 test forms

Effect	Form	Ability	1	2	3	4	5	6	7	8	9	10	11	12
Group	1011	1. Gates	1	**0.91**	**0.88**	0.65	0.72	0.69	-0.05	0.09	0.12	0.16	0.33	0.31
		2. WG		1	**0.90**	0.62	0.57	0.47	0.04	0.18	0.20	0.30	0.29	0.35
		3. Latent			1	0.53	0.66	0.71	-0.02	0.10	0.10	0.09	0.41	0.36
	1112	4. Gates				1	**0.93**	**0.90**	-0.47	-0.26	-0.29	0.09	0.10	0.21
		5. WG					1	**0.94**	-0.27	0.05	-0.04	0.17	0.24	0.33
		6. Latent						1	-0.27	0.03	-0.07	0.04	0.22	0.25
Interaction	1011	7. Gates							1	**0.88**	**0.93**	0.65	0.77	0.71
		8. WG								1	**0.98**	0.58	0.81	0.76
		9. Latent									1	0.66	0.84	0.79
	1112	10. Gates										1	**0.67**	**0.84**
		11. WG											1	**0.93**
		12. Latent												1

Note. WG = Word Generation; correlations of interest are highlighted in yellow.

Figure 3 contrasts the uniform and non-uniform DIF classifications within the two WG test forms and three ability estimates. The results suggest that there were few items (red dots in the middle) that showed no DIF across different test forms and ability estimates. The majority of items within the WG 1011 test form showed non-uniform DIF regardless of which ability estimate was used in the model. This pattern was less pronounced for the WG 1112 test form. Different findings for the two forms may stem from more limited power to detect statistically significant interactions in the 1112 test form because the the sample size was much smaller in the 2011-2012 cohort compared to the 2010-2011 cohort.

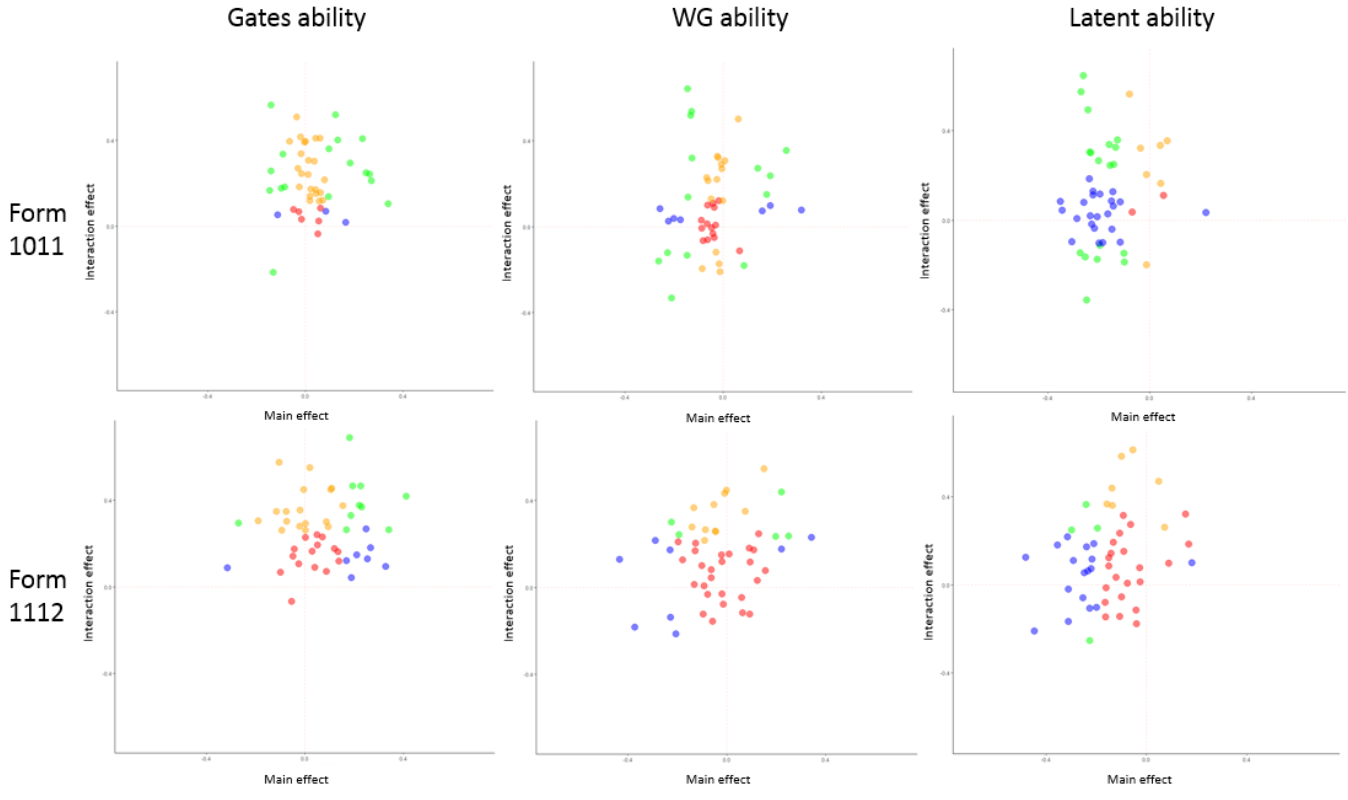


Figure 3. Scatterplots contrasting the uniform and non-uniform DIF classifications within the two WG test forms using three different ability estimates. Red dots represent items that do not show uniform or non-uniform DIF. Yellow dots represent items that do not show uniform DIF but show non-uniform DIF. Blue dots represent items that show uniform DIF but do not show non-uniform DIF. Green dots represent items that show both uniform and non-uniform DIF. ***To simplify, red dots represent items with no DIF, blue dots represent items with uniform DIF, while yellow and green dots represent items with non-uniform DIF.***

We have begun conducting DIF analyses using the IRT approach, which is preferable because it provides greater sensitivity and allows for analysis of items on both the 1011 and 1112 forms simultaneously using a common items equating approach to link the two forms due to the presence of 19 overlapping items. We are not yet ready to present those results in the report, but will complete these analyses over the summer during the final two months of the first project year. During the next reporting period we will report on DIF analyses using the IRT approach as well as the differential distractor functioning (DDF) analyses, which will also be completed over the summer and beginning of project year 2. After completing these analyses for DIF and DDF using the IRT approach, we will begin explanatory IRT analyses to examine why certain items are biased for ELs.

Opportunities for Training

Nothing to Report

How were the results communicated to Communities of Interest?

During the first reporting period there has not yet been anything of interest to a general audience of stakesholders. As we complete the database and the DIF, DDF, and e-IRT analyses, we epxect to develop a website for the project and plan to develop and submit both peer reviewed journal articles as well as submissions for paper presentations at professional society meetings. It is expected that the early work and releases will be more technical and generally of interest to modelers and measurement professionals, but as the work develops toward refining the assessment, the practical implications of the work will be more apparent and other kinds of dissemination activities and auidences will be possible.

II. PRODUCTS

No papers have been published to date and no presentations have been submitted to professional conferences. We anticipate one or more papers on DIF, DDF, and the e-IRT analyses in project year 2. Other products produced in project year 1 are the build out of the various databases of item characteristics. These separate, linked databases remain in development and are not yet ready for public release or dissemination as an integrated database on information on words, meanings, and test items.

What follows is a list of variables and descriptive statistics for all measures currently coded in the word meanings and features database. The database includes data on 6,221 meanings of 1,023 word forms. There are a total of 269 variables in our current database, and that does not include some important orthographic and phonological controls that we will also use in our analysis, along with features that pertain to the specific item, which involves the pairing of a target word and its meaning to a particular key word and set of distractors. This Q design matrix will be constructed from this word meanings and features database along with the other information that will be incorporated later.

Variable	Obs	Unique	Mean	Min	Max	Label
word	6221	1023	.	.	.	entry
numberofwo~s	6221	1	1	1	1	number of words
hyph	6221	2	.0011252	0	1	
synset	6221	5324	.	.	.	Wordnet. Meaning code.
sem_prec	6221	16	3.531747	0	17	BiFrost. Semantic Precision
poswn	6221	5	.	.	.	Wordnet:Part of speech
definition	6221	5316	.	.	.	Wordnet. Definition
syn	6221	72	7.247227	1	72	Wordnet. Ordinal synset number for each word. Note: Nouns first etc.
possyn	6221	59	4.825591	1	59	Wordnet. Ordinal synset number for each part of speech for a given word
pos	6221	4	.	.	.	Part of Speech
sem_pr~anpos	6221	205	3.531747	0	13	Wordnet. Mean of semantic precision for meanings of that POS
sem_prec_~an	6221	268	3.531747	0	13	Wordnet. Mean of semantic precision for meanings across POS
sem_prec_s~s	5869	263	1.235927	0	6.363961	Wordnet. SD of semantic precision for meanings of that POS
sem_prec_sd	6099	408	2.087838	0	7.071068	Wordnet. SD of semantic precision for meanings across POS
sem_pr~inpos	6221	14	2.195949	0	13	Wordnet. Min of semantic precision for meanings of that POS
sem_pre~xpos	6221	16	5.311204	0	17	Wordnet. Max of semantic precision for meanings of that POS
sem_prec_min	6221	13	1.13583	0	13	Wordnet. Min of semantic precision for meanings across POS
sem_prec_max	6221	16	6.966243	0	17	Wordnet. Max of semantic precision for meanings across POS
z_sem_prec	5273	26	-.1779605	-2.617673	5.239193	
z_sem_prec~n	5907	436	-.1856698	-2.122346	4.015063	Wordnet. Mean of Z score transformed semantic precision within POS
z_sem_prec~s	5273	260	-.1779605	-2.122346	4.015063	
pos_syn	6221	59	5.035525	1	59	
poly	6221	38	13.49445	1	72	Wordnet. Number of total meanings for a given word
pospoly	6221	32	9.07105	1	59	Wordnet. Numer of total meanings for a given word's specific pos
n_meanings	6221	19	3.77592	0	20	Wordnet: Noun meanings for that word
v_meanings	6221	29	7.979425	0	59	Wordnet: Verb meanings for that word

Variable	Obs	Unique	Mean	Min	Max	Label
j_meanings	6221	17	1.602315	0	27	Wordnet: Adjective meanings for that word
r_meanings	6221	8	.1367947	0	7	Wordnet: Adverb meanings for that word
orgword	6221	1	1	1	1	
hypernym1	5273	341	.	.	.	Hypernym 1
hypernym2	5273	845	.	.	.	Hypernym 2
hypernym3	4893	800	.	.	.	Hypernym 3
hypernym4	4084	594	.	.	.	Hypernym 4
hypernym5	3300	604	.	.	.	Hypernym 5
hypernym6	2744	750	.	.	.	Hypernym 6
hypernym7	2302	852	.	.	.	Hypernym 7
hypernym8	1837	747	.	.	.	Hypernym 8
hypernym9	1268	491	.	.	.	Hypernym 9
hypernym10	746	317	.	.	.	Hypernym 10
hypernym11	408	144	.	.	.	Hypernym 11
hypernym12	178	65	.	.	.	Hypernym 12
hypernym13	76	24	.	.	.	Hypernym 13
hypernym14	26	10	.	.	.	Hypernym 14
hypernym15	12	7	.	.	.	Hypernym 15
hypernym16	8	2	.	.	.	Hypernym 16
hypernym17	1	1	.	.	.	Hypernym 17
hypernym18	1	1	.	.	.	Hypernym 18
hypernym19	1	1	.	.	.	Hypernym 19
hypernym20	0	0	.	.	.	Hypernym 20
nhypernym1	5273	3882	18553.91	1	77304	Orninal count of words in the level 1 hypernym
thypernym1	5273	153	33983.8	1	77455	Total number of words in the level 1 hypernym
nhypernym2	5273	3251	8584.793	1	39359	Orninal count of words in the level 2 hypernym
thypernym2	5273	176	17035.87	2	39425	Total number of words in the level 2 hypernym
nhypernym3	5273	3369	4088.271	1	19300	Orninal count of words in the level 3 hypernym

Variable	Obs	Unique	Mean	Min	Max	Label
thypernym3	5273	146	8247.604	2	19302	Total number of words in the level 3 hypernym
nhypernym4	5273	3708	5456.549	1	21986	Orninal count of words in the level 4 hypernym
thypernym4	5273	146	10882.58	1	21994	Total number of words in the level 4 hypernym
nhypernym5	5273	3719	9384.661	1	41484	Orninal count of words in the level 5 hypernym
thypernym5	5273	179	18681.48	1	41484	Total number of words in the level 5 hypernym
nhypernym6	5273	3705	14013.69	1	55570	Orninal count of words in the level 6 hypernym
thypernym6	5273	183	27985.98	1	55573	Total number of words in the level 6 hypernym
nhypernym7	5273	3836	19438.49	1	66268	Orninal count of words in the level 7 hypernym
thypernym7	5273	158	38638.51	1	66360	Total number of words in the level 7 hypernym
dup_all	5273	2	.0003793	0	1	
xhap	6221	1	0	0	0	xhper: Hapaxes
DaviesID	5926	1426	8785.355	49	98706	Davies: rank order 1-100,000
l1	5926	1096	.	.	.	Davies: lemma
pos1	5926	11	.	.	.	Davies: part of speech
caps	5926	57	.	.	.	Davies: Percent of tokens that are capitalized.
usuk	5926	2	.	.	.	Davies: US spelling or UK spelling
freq	5926	1354	27608.3	17	1037828	Davies: Raw frequency (# tokens) in the 450 million word Corpus of Contemporary
coca	5926	1127	59.45847	.04	2235.11	Davies: Frequency (per million words) in the 450 million word Corpus of Contempo
bnc	5926	1122	57.10914	0	1739.18	Davies: Frequency (per million words) in the 100 million word British National C
soap	5926	777	61.60879	0	2643.76	Davies: Frequency (per million words) in the 100 million word Corpus of American
yr195089	5926	1101	60.7571	0	3614.58	Davies: Frequency (per million words) in the Corpus of Historical American Engli
yr190049	5926	1077	57.22091	0	2595.89	Davies: Frequency (per million words) in the Corpus of Historical American Engli
yr1800s	5926	1033	51.99884	0	1915.24	Davies: Frequency (per million words) in the Corpus of Historical American Engli
coca_spok	5926	966	57.34772	0	1673.43	Davies: Frequency (per million words) in COCA genres: spoken
coca_fic	5926	1004	62.90311	0	4080.19	Davies: Frequency (per million words) in COCA genres: fiction
coca_mag	5926	1122	56.80854	.03	779.7	Davies: Frequency (per million words) in COCA genres: popular magazines

Variable	Obs	Unique	Mean	Min	Max	Label
coca_news	5926	1076	63.00948	0	4604.43	Davies: Frequency (per million words) in COCA genres: newspapers
coca_acad	5926	1082	57.46912	0	996.49	Davies: Frequency (per million words) in COCA genres: academic journals
bnc_spok	5926	457	60.34234	0	2524.77	Davies: Frequency (per million words) in BNC genres: spoken
bnc_fic	5926	580	63.84085	0	5072.25	Davies: Frequency (per million words) in BNC genres: fiction
bnc_mag	5926	490	55.66514	0	820.57	Davies: Frequency (per million words) in BNC genres: popular magazines
bnc_news	5926	556	60.33756	0	3766.14	Davies: Frequency (per million words) in BNC genres: newspapers
bnc_noac	5926	695	55.50003	0	1152.94	Davies: Frequency (per million words) in BNC genres: non-academic journals
bnc_acad	5926	672	57.78759	0	777.15	Davies: Frequency (per million words) in BNC genres: academic journals
bnc_misc	5926	766	60.32084	0	695.41	Davies: Frequency (per million words) in BNC genres: miscellaneous
perc_coca	5926	45	.0708809	0	.57	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: COC
perc_bnc	5926	87	.2701822	0	.89	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: BNC
perc_soap	5926	81	.1143554	0	1	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: SOA
perc_195089	5926	30	.0465305	0	.55	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: 2)
perc_190049	5926	32	.0531269	0	.47	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: 2)
perc_1800s	5926	58	.1380324	0	.61	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: 2)
perc_coca_~k	5926	55	.0683766	0	.76	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: COC
perc_coca_~c	5926	62	.1160867	.01	.84	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: COC
perc_coca_~g	5926	42	.0610175	0	.54	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: COC
perc_coca_~s	5926	41	.0574519	0	.72	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: COC
perc_coca_~d	5926	59	.1041056	0	.74	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: COC
perc_bnc_s~k	5926	73	.151922	.01	.89	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: BNC
perc_bnc_fic	5926	99	.3673338	0	.99	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: BNC
perc_bnc_mag	5926	101	.4067769	0	1	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: BNC
perc_bnc_n~s	5926	80	.2428755	0	.9	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: BNC
perc_bnc_n~c	5926	96	.3349241	0	.96	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: BNC
perc_bnc_a~d	5926	96	.305162	0	.96	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: BNC
perc_bnc_m~c	5926	87	.2753426	0	.88	Davies: Percentage of texts (0.00-1.00) that contain the word at least once: BNC
raw_bnc	5926	1122	5710.914	0	173918	Davies: Raw token frequency in BNC

Variable	Obs	Unique	Mean	Min	Max	Label
raw_soap	5926	777	6160.879	0	264376	Davies: Raw token frequency in SOAP
raw_195089	5926	1101	5933.119	0	352975	Davies: Raw token frequency in COHA: 1950-89
raw_190049	5926	1140	6832.32	0	309956	Davies: Raw token frequency in COHA: 1900-49
raw_1800s	5926	1092	6808.567	0	250776	Davies: Raw token frequency in COHA: 1800s
raw_coca_s~k	5926	966	5480.45	0	159921	Davies: Raw token frequency in COCA genre: spoken
raw_coca_fic	5926	1004	5688.268	0	368969	Davies: Raw token frequency in COCA genre: fiction
raw_coca_mag	5926	1122	5428.565	3	74507	Davies: Raw token frequency in COCA genre: popular magazines
raw_coca_n~s	5926	1076	5779.083	0	422307	Davies: Raw token frequency in COCA genre: newspapers
raw_coca_a~d	5926	1082	5233.504	0	90747	Davies: Raw token frequency in COCA genre: academic journals
raw_bnc_spok	5926	457	601.2335	0	25156	Davies: Raw token frequency in BNC genre: spoken
raw_bnc_fic	5926	580	1015.662	0	80696	Davies: Raw token frequency in BNC genre: fiction
raw_bnc_mag	5926	490	404.2401	0	5959	Davies: Raw token frequency in BNC genre: popular magazines
raw_bnc_news	5926	556	631.5174	0	39418	Davies: Raw token frequency in BNC genre: newspapers
raw_bnc_noac	5926	695	915.4865	0	19018	Davies: Raw token frequency in BNC genre: non-academic journals
raw_bnc_acad	5926	672	885.9792	0	11915	Davies: Raw token frequency in BNC genre: academic journals
raw_bnc_misc	5926	766	1256.795	0	14489	Davies: Raw token frequency in BNC genre: miscellaneous
raw_txt_coca	5926	1308	13457.35	16	108548	Davies: Raw number of texts in COCA
raw_txt_bnc	5926	919	1094.838	0	3608	Davies: Raw number of texts in BNC
raw_txt_soap	5926	764	2517.453	0	21945	Davies: Raw number of texts in SOAP
raw_t~195089	5926	1011	1992.67	0	23409	Davies: Raw number of texts in COHA: 1950-89
raw_t~190049	5926	1029	2107.343	0	18767	Davies: Raw number of texts in COHA: 1900-49
raw_tx~1800s	5926	972	1504.629	0	6647	Davies: Raw number of texts in COHA: 1800s
raw_txt_co~k	5926	911	2632.039	0	29060	Davies: Raw number of texts in COCA genre: spoken
raw_txt_co~c	5926	937	2066.323	0	16162	Davies: Raw number of texts in COCA genre: fiction
raw_txt_co~g	5926	1069	3255.095	3	28994	Davies: Raw number of texts in COCA genre: popular magazines
raw_txt_co~s	5926	1022	3287.938	0	40885	Davies: Raw number of texts in COCA genre: newspapers
raw_txt_co~d	5926	982	2215.96	0	15658	Davies: Raw number of texts in COCA genre: academic journals
raw_txt_bn~k	5926	307	128.9615	0	800	Davies: Raw number of texts in BNC genre: spoken

Variable	Obs	Unique	Mean	Min	Max	Label
raw_txt_b~ic	5926	360	170.4165	0	461	Davies: Raw number of texts in BNC genre: fiction
raw_txt_bn~g	5926	203	85.8542	0	210	Davies: Raw number of texts in BNC genre: popular magazines
raw_txt_bn~s	5926	306	125.6993	0	466	Davies: Raw number of texts in BNC genre: newspapers
raw_txt_b~ac	5926	409	178.9026	0	514	Davies: Raw number of texts in BNC genre: non-academic journals
raw_txt_bn~d	5926	363	152.8296	0	480	Davies: Raw number of texts in BNC genre: academic journals
raw_txt_b~sc	5926	485	252.1745	0	803	Davies: Raw number of texts in BNC genre: miscellaneous
cocatotal	5926	1138	64.06657	.04	2371.24	Total number of meanings in COCA after collapsing POS
dup	5926	1	0	0	0	
mergedavies	6221	2	2.90516	1	3	
occurtotal	5648	13	33.91802	18	1934	Kuperman: total number of times the word occurs in the trimmed data
occurnum	5648	19	33.85127	13	1930	Kuperman: number of responders who gave numeric ratings to the word, rather than
aoamean	5648	439	.	.	.	Kuperman: mean age of acquisition rating (in years of age)
aoasd	5648	274	.	.	.	Kuperman: standard deviation of age of acquisition rating
_mergeAOA	6221	2	2.815785	1	3	
aoahap	6221	1	0	0	0	Kuperman: AOA Hapaxes
mean_cos	6096	970	.0165629	.0045543	.3776976	Hoffman: Mean cosine similarity between contexts containing the word
semd	6096	970	1.875406	.4228558	2.341577	Hoffman: Semantic Diversity
_mergesemd	6221	2	2.959814	1	3	
semdhap	6221	1	0	0	0	Hoffman: SemD Hapaxes
cd	6206	531	844.0079	1	10864	Adelman: Contextual diversity
_mergecont	6221	2	2.995178	1	3	
cdhap	6221	1	0	0	0	Adelman: CD Hapaxes
sfi	6220	326	54.46571	11.5	73.8	Zeno:Standardized frequency index
d	6220	929	.8026088	0	.9945	Zeno:Dispersion
u	6220	308	95.57337	.0014	2382	Zeno: U
f	6220	653	1833.448	1	58789	Zeno: f
gr1	3823	151	212.1263	0	17700	Zeno: gr1
gr2	5237	165	124.4726	0	8732	Zeno: gr2

Variable	Obs	Unique	Mean	Min	Max	Label
gr3	5677	162	107.4844	0	5894	Zeno: gr3
gr4	5828	177	103.5287	0	4514	Zeno: gr4
gr5	5904	178	97.75745	0	2570	Zeno: gr5
gr6	5909	192	95.85734	0	1877	Zeno: gr6
gr7	5912	186	95.45974	0	1586	Zeno: gr7
gr8	5912	193	94.90629	0	1501	Zeno: gr8
gr9	5913	194	94.36513	0	1507	Zeno: gr9
gr10	5917	196	92.99983	0	1498	Zeno: gr10
gr11	5917	202	92.6892	0	1514	Zeno: gr11
gr12	5920	207	91.93733	0	1460	Zeno: gr12
gr13	5915	222	90.38664	0	1268	Zeno: gr13
_mergezeno	6221	2	2.999679	1	3	
sub_id	6157	301	449.2186	21	759	English Lex: subject ID for naming or lexical decision experiments
trial	6157	857	1697.014	3	3371	English Lex: position a trial appears within an experimental block
type	6157	1	1	1	1	English Lex word (1) or non-word (0)
d_accuracy	6157	2	.9803476	0	1	English Lex: trial accuracy
d_rt	6157	522	662.8809	26	4000	English Lex: trial response time (in msec)
d_zscore	6157	991	-.4808429	-2.541589	8.530417	English Lex: z-standardized response times
d_rt_mean	6157	969	654.8235	448.1667	1413.071	English Lex: mean of response time across participants and trials for the word
d_accuracy~n	6157	45	.9981601	.3076923	4.75	English Lex: mean of trial accuracy across participants and trials for the word
_mergeEngLex	6221	2	2.979425	1	3	
freqcount	6161	685	5605.983	1	168631	SUBTLEX: number of times the word appears in the SUBTLEX-US corpus of 51 million
cdcount	6161	650	1662.154	1	8273	SUBTLEX: number of films in which the word appears
freqlow	6161	661	5020.878	1	160730	SUBTLEX: number of times the word appears in the corpus starting with a lowercas
cdlow	6161	623	1588.929	1	8267	SUBTLEX: number of films in which the word appears starting with a lowercase let

Variable	Obs	Unique	Mean	Min	Max	Label
subtlwf	6161	685	109.9212	.0196078	3306.49	SUBTLEX: word frequency per million words
lg10wf	6161	685	2.962733	.30103	5.22694	SUBTLEX: log10(FREQcount+1) and has 4-digit precision
subtlcd	6161	650	19.81585	.0119218	98.62899	SUBTLEX: percentage of films the word appears
lg10cd	6161	650	2.762626	.30103	3.917716	SUBTLEX: log10(CDcount + 1) and has 4-digit precision
_mergesubt	6221	2	2.98071	1	3	
vmeansum	4185	316	5.509548	1.62	8	Warriner: Overall mean valence
vsdsum	4185	147	1.62227	.61	2.63	Warriner: standard deviation of overall valence
vratsum	4185	21	24.24134	18	870	Warriner: number of contributing ratings for valence
ameansum	4185	248	4.126076	1.95	7.24	Warriner: Overall mean arousal
asdsum	4185	135	2.280368	1.36	3.27	Warriner: standard deviation of overall arousal
aratsum	4185	26	23.77515	17	917	Warriner: number of contributing ratings for arousal
dmeansum	4185	274	5.501553	2.43	7.63	Warriner: Overall mean dominance
dsdsum	4185	140	2.12536	1.04	3.05	Warriner: standard deviation of overall dominance
dratsum	4185	41	24.30968	15	978	Warriner: number of contributing ratings for dominance
vmeanm	4185	213	5.446996	1.2	8.6	Warriner: mean valence: male
vsdm	4185	191	1.605245	0	3.7	Warriner: standard deviation of valence: male
vratm	4185	25	9.044205	2	335	Warriner: number of contributing ratings for valence: male
vmeanf	4185	288	5.536249	1.47	8.38	Warriner: mean valence: female
vsdf	4185	178	1.575037	.63	3.03	Warriner: standard deviation of valence: female
vratf	4185	28	14.86906	6	529	Warriner: number of contributing ratings for valence: female
ameanm	4185	186	4.139422	1.8	7.71	Warriner: mean arousal: male
asdm	4185	181	2.281952	.49	3.53	Warriner: standard deviation of arousal: male
aratm	4185	22	8.628674	4	344	Warriner: number of contributing ratings for arousal: male
ameanf	4185	241	4.098057	2	7.33	Warriner: mean arousal: female
asdf	4185	170	2.258643	1.25	3.53	Warriner: standard deviation of arousal: female
aratf	4185	27	15.01792	8	566	Warriner: number of contributing ratings for arousal: female
dmeanm	4185	215	5.453379	2.5	7.71	Warriner: mean dominance: male
dsdm	4185	176	2.107957	.98	3.63	Warriner: standard deviation of dominance: male

Variable	Obs	Unique	Mean	Min	Max	Label
dratm	4185	29	11.0638	5	431	Warriner: number of contributing ratings for dominance: male
dmeanf	4185	253	5.561699	1.5	8.3	Warriner: mean dominance: female
dsdf	4185	191	2.109646	.73	3.39	Warriner: standard deviation of dominance: female
dratf	4185	33	13.0282	6	542	Warriner: number of contributing ratings for dominance: female
vmeany	4185	258	5.498215	1	8.3	Warriner: mean valence: younger
vsdy	4185	184	1.641481	0	3.2	Warriner: standard deviation of valence: younger
vraty	4185	25	12.43417	4	416	Warriner: number of contributing ratings for valence: younger
vmeano	4185	272	5.55248	1.58	8.29	Warriner: mean valence: older
vsdo	4185	185	1.540827	.42	3.21	Warriner: standard deviation of valence: older
vrato	4185	23	11.43154	5	447	Warriner: number of contributing ratings for valence: older
ameany	4185	215	4.164631	1.67	7.29	Warriner: mean arousal: younger
asdy	4185	189	2.327309	.95	3.77	Warriner: standard deviation of arousal: younger
araty	4185	27	10.64946	5	414	Warriner: number of contributing ratings for arousal: younger
ameano	4185	233	4.087168	2	7.45	Warriner: mean arousal: older
asdo	4185	178	2.218645	1.1	3.47	Warriner: standard deviation of arousal: older
arato	4185	26	13.05974	7	499	Warriner: number of contributing ratings for arousal: older
dmeany	4185	258	5.494155	2.42	7.73	Warriner: mean dominance: younger
dsdy	4185	176	2.216889	.6	3.76	Warriner: standard deviation of dominance: younger
draty	4185	36	13.46141	5	537	Warriner: number of contributing ratings for dominance: younger
dmeano	4185	246	5.460045	1.75	8.44	Warriner: mean dominance: older
dsdo	4185	205	1.907716	.53	3.42	Warriner: standard deviation of dominance: older
drato	4185	27	10.76344	4	437	Warriner: number of contributing ratings for dominance: older
vmeanl	4185	264	5.449364	1.6	8.38	Warriner: mean valence: low education
vsdl	4185	173	1.627192	.64	2.87	Warriner: standard deviation of valence: low education
vratl	4185	23	13.80048	6	478	Warriner: number of contributing ratings for valence: low education
vmeanh	4185	244	5.613059	1.64	8.5	Warriner: mean valence: high education
vsdh	4185	182	1.534769	.35	2.92	Warriner: standard deviation of valence: high education
vrath	4185	23	10.44086	6	392	Warriner: number of contributing ratings for valence: high education

Variable	Obs	Unique	Mean	Min	Max	Label
ameanl	4185	238	4.105235	1.83	7.8	Warriner: mean arousal: low education
asdl	4185	187	2.398502	1.11	3.92	Warriner: standard deviation of arousal: low education
aratl	4185	30	11.98256	6	509	Warriner: number of contributing ratings for arousal: low education
ameanh	4185	229	4.207233	1.5	7.62	Warriner: mean arousal: high education
asdh	4185	192	.	.	.	Warriner: standard deviation of arousal: high education
arath	4185	27	11.79259	3	408	Warriner: number of contributing ratings for arousal: high education
dmeanl	4185	261	5.600198	2	8.43	Warriner: mean dominance: low education
dsdl	4185	186	2.27563	.79	3.5	Warriner: standard deviation of dominance: low education
dratl	4185	33	13.02007	6	533	Warriner: number of contributing ratings for dominance: low education
dmeanh	4185	235	5.391701	1.43	7.94	Warriner: mean dominance: high education
dsdh	4185	196	1.874858	.52	4.36	Warriner: standard deviation of dominance: high education
drath	4185	28	11.28961	3	445	Warriner: number of contributing ratings for dominance: high education
_mergevale~e	6221	2	2.345443	1	3	
bigram	3820	1	0	0	0	Brysbaert Concreteness: Indicates if stimuli is a bigram
concm	3820	276	3.122772	1.25	5	Brysbaert Concreteness: Mean concreteness rating
concsd	3820	129	1.206448	0	1.75	Brysbaert Concreteness: SD concreteness rating
unknown	3820	7	.1780105	0	127	Brysbaert Concreteness: Number who did not know word
total	3820	13	52.87225	23	6064	Brysbaert Concreteness: Total Participants
percent_kn~n	3820	12	.9966544	.85	1	Brysbaert Concreteness: Percentage that knew word
_mergeconc~e	6221	2	2.228098	1	3	
imag	1256	138	421.3065	246	639	Bird: imageability rating
_mergeimag	6221	2	1.403794	1	3	
dupword	6221	73	7.227616	0	72	
dupwordpos	6221	60	4.978942	0	59	

III. Participants & Other Collaborating Organizations

KEY PERSONNEL

Name:	Dr. David J. Francis
Effort	1 month
Role	PI/PD
Location	TIMES, University of Houston, Houston, TX, United States
Description	Directs the overall program of research, provides guidance on statistical modeling and analyses, coding, study design and analysis. Overall responsibility for budget, managing and approving expenditures, and maintaining study timelines and communication with IES. (Dr. Francis effort is paid in the summer, which is not included in this initial reporting period and thus his funding is not reflected in the SF 424 Expenditure Report).

Name:	Dr. Paulina A. Kulesz
Effort	2.6 months
Role	Co-Investigator
Location	TIMES, University of Houston, Houston, TX, United States
Description	Dr. Kulesz is responsible for carrying out the analyses associated with DIF, DDF, and e-IRT.

Name:	Dr. Autumn McIlraith
Effort	9 months
Role	Post-Doctoral Fellow
Location	TIMES, University of Houston, Houston, TX, United States
Description	Dr. McIlraith is a post-doctoral fellow working on the project and is based at TIMES at the University of Houston. She works with Drs. Francis, Kulesz, and Lawrence on the coding of item features, especially those characteristics of a linguistic or orthographic nature, such as LSA, LD, and OLD20, and is asisting in the design, organization, and compilation of the database, as well as in the design and construction of tables and graphical displays of results in order to communicate findings as clearly as possible. Dr. McIlraith will also assist Drs. Francis and Kulesz in the DIF, DDF, and e-IRT analyses.

Name:	Dr. Joshua Lawrence
Effort	1 month
Role	Co-Investigator
Location	Lawrence Consulting, MA, United States

Description	Dr. Lawrence has directed the collection and coding of information on words, word meanings, and items. Dr. Lawrence has taken the lead on the identification of key features to code for words and items as well as directing the design of the database and has directed the extraction of information from the relevant databases that are publicly available. He is also the lead contact with the state database that contains information on student characteristics and is a member of the team that developed the Word Generation Academic Vocabulary Assessment. He and Dr. Francis collaborated on the original Word Generation Efficacy Trial that was funded by IES and that is providing the student data for the current project in Phase I. Dr. Lawrence's effort is contracted through Lawrence Consulting. He is also a faculty member at the University of Oslo and spends the majority of the calendar year in residence in Oslo, Norway. He does spend some extended time each calendar year in the US. The team communicates weekly with Dr. Lawrence via email and through a weekly conference call hosted on Webex/Zoom. He regularly attends these meetings, regardless of his location.

OTHER PERSONNEL

Name:	Martin Walczak
Effort	25% effort for three months
Role	Graduate Student Research Assistant
Location	TIMES, University of Houston, Houston, TX, United States
Description	Mr. Walczak worked with Drs. Kulesz and McIlraith on the coding of words based on the number of definitions they have. He has no access to student data.
Name:	Katharina Roittner
Effort	10% effort for 2 months
Role	Graduate Student Research Assistant, MA Student in Educational Psychology
Location	University of Oslo, Oslo, Norway
Description	Ms. Roittner worked with Drs. Lawrence and McIlraith on the coding of words based on the number of definitions they have. She has access to student data.

OTHER SUPPORT FOR KEY PERSONNEL

There have been no changes to the other support for the PD/PI or other Key Personnel since the Just in Time information was submitted.

PARTICIPATING ORGANIZATIONS

University of Houston, Houston, TX

IV. IMPACT

The project has had no measureable impact to date as we are still in data compilation and analysis mode in Phase I of the work.

V. CHANGES/PROBLEMS

No significant problems were encountered in Project Year 1. We have been meeting successfully every week as a team via Webex/Zoom and making good progress on our objectives for Year 1. We feel that the project is on track as currently being implemented. For year 2, we have decided not to hire a second post-doc. Rather, we propose to increase the effort of Dr. Lawrence to one day per week.

VI. SPECIAL REPORTING REQUIREMENTS

No special reporting requirements were listed in the Notice of Grant Award.

VII. BUDGETARY INFORMATION

See attached SF 424 for the Expenditure Report from the Reporting Period

BUDGET JUSTIFICATION FOR GRANT YEAR 2 ACTIVITIES

Key Personnel:
David J. Francis serves as the Principal Investigator on the project. He will direct and oversee the project, statistical analyses related to the project aims, and statistical reporting. In particular, he will supervise the work of Dr. Kulesz who will implement the explanatory item response models, differential item functioning, and differential distractor functioning analyses, and Dr. Autumn McIlraith, post-doctoral Fellow working on the project. Funds are requested to cover one month (8%) of his summer salary and fringe in each year of the project.

Paulina A. Kulesz serves as co-Investigator on the project. She will work with Dr. Francis to oversee any data management tasks that are necessary to prepare the data for analysis. Dr. Kulesz will be primarily responsible for implementing the explanatory item response models, differential item functioning, and differential distractor functioning analyses. She will also assist Drs. Francis and Lawrence in scientific reporting. Funds are requested to cover 4.8 months (40%) of her salary and fringe in each year of the project.

Dr. Joshua Lawrence, Lawrence Consulting and the University of Oslo, serves as co-Investigator. Dr. Lawrence's role on the project is described more fully under Consultant Services, reflecting how he will be paid on the project. However, his role is that of co-Investigator, reflecting his central role in the proposal. When the project was originally submitted, Dr. Lawrence was at the University of California at Irvine. Since then he has moved to the University of Oslo, thus making it necessary to engage Dr. Lawrence through a consulting services contract while serving as a co-Investigator. His effort in year 2 of the project will be 2 calendar months.

Dr. Autumn McIlraith, Post-doctoral Fellow, was hired in year 1 of the project as a post-doctoral Fellow at TIMES, University of Houston. She is employed full-time on the project in year 2 to assist Dr. Kulesz with statistical analyses in the second year of the project.

Other Personnel:

Laudemer Vigilia, Data manager, will be responsible for organizing, controlling, and aggregating data. Funds are requested to cover 16.67% effort (2 calendar months) in year one (outside of the first reporting period) and two.

Project Manager 1 (to be determined) at 8% effort (1 calendar month) in year two. This person will be in charge of examiner training and the coordination of data collection at the schools. This person will maintain a data collection management system that will organize and register the staff, materials and sites involved in data collection. In Year 2, only pilot testing of new items is proposed for new data collection, so it is not anticipated that we will hire and train examiners. Rather, data collection will be handled by the project manager and student volunteers.

Fringe Benefits:
Fringe benefits are based on actual amounts and calculated using a Fringe Benefits Calculator supplied by UH Office of Contracts and Grants.

Travel:
Funds are requested to support travel by Drs. Francis and/or Kulesz to the required IES meetings in year 2, and for key personnel to present findings from year 1 activities at professional conferences. In terms of travel to the IES meetings, the estimated cost per person per trip will be $1,367 in year one and will include airfare, lodging, ground transportation, meals and incidentals. In terms of travel to a conference (including escalation in each year), the estimated cost per person per trip will be $2,465 and will include airfare, lodging, ground transportation, meals and incidentals, and conference registration fees. Funds are also requested to support travel to Houston for the 4 consultants at $3,832. Costswill include airfare, lodging, ground transportation, meals and incidentals. Please see below the breakdown for travel for all years of the grant:

PI travel to IES Meetings

Destination	# Days	# Traveler	Airfare	Tot. Per Diem	Total Travel	
DC	2	1	$ 675.00	$ 346.00	$ 1,367.00	year 1
DC	2	1	$ 700.00	$ 354.00	$ 1,408.00	year 2
DC	2	1	$ 725.00	$ 362.50	$ 1,450.00	year 3
DC	2	1	$ 750.00	$ 372.00	$ 1,494.00	year 4
					$ 5,719.00	

Conference Travel

Destination	# Days	# Traveler	Airfare	Tot. Per Diem	Registration	Total Travel	
TBD	4	1	$ 650.00	$ 353.75	$ 400.00	$ 2,465.00	year 2
TBD	4	1	$ 650.00	$ 353.75	$ 400.00	$ 2,465.00	year 3
TBD	4	1	$ 650.00	$ 353.75	$ 400.00	$ 2,465.00	year 4
						$ 7,395.00	

Consultant travel to Houston

Destination	# Days	# Traveler	Airfare	Tot. Per Diem		Total Travel	
Houston	2	4	$ 500.00	$ 229.00		$ 3,832.00	year 1
Houston	2	4	$ 500.00	$ 229.00		$ 3,832.00	year 2
Houston	2	4	$ 500.00	$ 229.00		$ 3,832.00	year 3
Houston	2	4	$ 500.00	$ 229.00		$ 3,832.00	year 4
						$ 15,328.00	

Other Direct Costs:

Consultants Services: Funds are requested to cover a consultant rate of $1000 per day per consultant in each year of the project. The consultants are Drs. Mikyung Wolf, Catherine Snow, Young-Suk Kim, and Paul DeBoeck. We are budgeting to cover the consultant rate for 4 consultants who will consult on the content, substantive nature of obtained findings, and statistical analyses. Depending on the needs, we are proposing to meet with the consultants 1 to 2 days per year.

Dr. Joshua Lawrence, Co-Investigator: Funds are requested for Dr. Josh Lawrence. Dr. Lawrence is listed as a consultant because of his relocation from UC Irvine to the University of Oslo during the period of time in between submission of the proposal and its being awarded. Dr. Lawrence will commit two months of effort to the project in year 2. Dr. Lawrence will serve as co-Investigator on the overall project, will work with Dr. Francis in directing all aspects of the work conducted, including development of the item coding for existing items, integration of new language information regarding existing items, and development of item models DIF analyses, and DIF distractor analyses, planning and conducting the pilot data collection, statistical analyses of pilot data, and the development of new items, and leading the preparation of presentations, reports, and publications about project findings.

Indirect Costs:
Indirect costs are calculated on a modified total direct costs basis using the DHHS-approved rate of 50.5%.