Electronic Supplementary Material (ESI) for Chemistry Education Research and Practice

Assessing the relation between language comprehension and performance in general chemistry

Daniel T. Pyburn[a], Samuel Pazicni*[a], Victor A. Benassi[b], and Elizabeth E. Tappin[c]

[a] *Department of Chemistry, University of New Hampshire, Durham, New Hampshire, USA. E-mail: sam.pazicni@unh.edu*
[b] *Department of Psychology and Center for Excellence in Teaching and Learning, University of New Hampshire, Durham, New Hampshire, USA.*
[c] *Center for Excellence in Teaching and Learning, University of New Hampshire, Durham, New Hampshire, USA*

Appendices

Contents

Appendix 1. Data screening and descriptive statistics for data collected in all General Chemistry courses.

Chem A. Data were collected from a total of 1334 students. SAT data were missing for 226 students; ACS Toledo Placement Exam data were missing for 96 students; and ACS First Term General Chemistry Paired Questions Exam data were missing for 379 students. Two univariate outliers ($p < .001$) were identified in each of the SAT-CR, SAT-Math, and ACS Toledo Exam score distributions; each of these outliers corresponded to a different student. All univariate outliers were filtered from corresponding analyses. The criterion for multivariate outliers was Mahalanobis distance at $p < .001$ (Tabachnick and Fidell, 2013, pp 99); no mulivariate outliers were identified. Once outliers were filtered, significance testing revealed some divergence from normality in the SAT-CR, ACS Toledo Exam, and ACS First Term General Chemistry Paired Questions Exam data. Upon examination it was determined that the skewness and kurtosis values for these data were within the range of ±1. The inferential statistics used in this study were robust to these modest violations of the normality assumption (Cohen *et al*., 2003, pp 41). Descriptive statistics for the screened Chem A data used in reported analyses are provided in Table A1.

Table A1. Descriptive statistics for predictor/outcome variables in Chem A following removal of univariate and multivariate outliers.

	SAT-Critical Reading	*SAT-Math*	*ACS Toledo Placement Exam*	*ACS First Term General Chemistry Paired Questions Exam*
N	1126	1126	1236	955
Mean	553.32	564.09	47.68%	63.28%
Standard Deviation	76.20	71.54	12.33	16.23
Skewness	.241 (Std. error = .073)	–.028 (Std. error = .073)	.342 (Std. error = .070)	–.132 (Std. error = .079)
Excess Kurtosis	–.110 (Std. error = .146)	–.268 (Std. error = .146)	–.053 (Std. error = .193)	–.579 (Std. error = .158)

Chem B. Data were collected from a total of 579 students. SAT data were missing for 82 students; ACS General Chemistry Conceptual Exam (First Term) data were missing for 60 students; and ACS Special Exam (1997) data were missing for 12 students. No univariate or multivariate outliers were identified. Significance testing revealed some divergence from normality in the SAT-CR, ACS General Chemistry Conceptual Exam (First Term), and ACS Special Exam (1997) Exam data. Upon examination it was determined that the skewness and kurtosis values for these data were within the range of ±1. The inferential statistics used in this study are robust to these modest violations of the normality assumption (Cohen *et al*., 2003, pp 41). Descriptive statistics for the screened Chem B data used in reported analyses are provided in Table A2.

Table A2. Descriptive statistics for predictor/outcome variables in Chem B following removal of univariate and multivariate outliers.

	SAT-Critical Reading	*SAT-Math*	*ACS General Chemistry Conceptual Exam (First Term)*	*ACS Special Exam (1997)*
N	497	497	519	567
Mean	559.22	575.37	47.47%	64.83%
Standard Deviation	78.95	74.21	15.41	18.55
Skewness	.374 (Std. error = .110)	.066 (Std. error = .110)	.484 (Std. error = .107)	.028 (Std. error = .103)
Excess Kurtosis	–.010 (Std. error = .219)	–.156 (Std. error = .219)	–.170 (Std. error = .214)	–.555 (Std. error = .205)

Chem C. Data were collected from a total of 595 students. 7 of these students had been enrolled in Chem C twice; the most recent record for each of these students was retained in the data set. SAT data were missing for 57 students; Gates-MacGinitie Reading Test data were missing for 56 students; ACS Toledo Placement Exam data were missing for 42 students; and ACS General Chemistry (Conceptual) Exam data were missing for 16 students. Midterm exam data were missing in 33 cases involving 23 students (summarized in Table A3). One univariate outlier ($p < 0.001$) was identified in each of the SAT-CR, SAT-Math, ACS Toledo Exam, ACS General Chemistry (Conceptual) Exam score distributions; each of these outliers corresponded to a different student. Seven univariate outliers involving four students were identified from standardized midterm exam score distributions. All univariate outliers were filtered from corresponding analyses. Mulivariate outliers were identified via Mahalanobis distances at $p < .001$. Four multivariate outliers were found involving ACS Exam data, while two were found for midterm exam data; each of these outliers corresponded to the same student who had an uncharacteristically low ACS Toledo Placement Exam score. This student's data was removed from analyses relating prior knowledge to course performance. Once outliers were filtered, significance testing revealed some divergence from normality in the SAT-CR and midterm exam data. Upon examination it was determined that the skewness and kurtosis values for these data were within the range of ±1. The inferential statistics used in this study are robust to these modest violations of the normality assumption (Cohen *et al*., 2003, pp 41). Descriptive statistics for the screened Chem C data used in reported analyses are provided in Table A4.

Table A3. Summary of missing midterm exam data for Chem C.

Description	*Frequency*	*Percent*
Completed all exams	572	96%
Completed three of four exams	16	2.7%
Completed two of four exams	4	0.7%
Completed one of four exams	3	0.5%

Table A4. Descriptive statistics for predictor/outcome variables in Chem C following removal of univariate and multivariate outliers.

	SAT-Critical Reading	*Gates-MacGinitie Reading Test*	*SAT-Math*	*ACS Toledo Placement Exam*	*ACS General Chemistry (Conceptual) Exam*
N	537	538	537	553	578
Mean	557.73	67.86%	620.47	50.69%	64.65%
Standard Deviation	72.33	15.88	66.63	12.82	12.83
Skewness	.280 (Std. error = 0.105)	–.135 (Std. error = 0.105)	–.063 (Std. error = 0.105)	.003 (Std. error = 0.104)	–.043 (Std. error = 0.102)
Excess Kurtosis	.066 (Std. error =.210)	–.301 (Std. error = .210)	–.053 (Std. error = .210)	–.005 (Std. error = .207)	–.224 (Std. error = .203)
	Standardized Midterm Exam Scores				
N	2430				
Mean	0.011				
Standard Deviation	0.978				
Skewness	–.421 (Std. error = 0.051)				
Excess Kurtosis	–.234 (Std. error = 0.101)				

Appendix 2. Correlations of performance predictors and ACS exam for all General Chemistry courses.

Two-tailed Pearson correlations were calculated to establish relationships between each of three predictor variables (prior chemistry knowledge, language comprehension ability, and math ability) and course performance as measured by ACS exams. This was done to explore the potential for multicollinearity in our HLMs and regression models. The relative magnitudes of Pearson correlations were interpreted using the qualitative guidelines described by Cohen (Cohen, 1988): small (.10), medium (.30) and large (.50).

Correlations of comprehension ability and ACS exam score in all courses were statistically significant with medium effect sizes (r = +.35 - +.45). Of note, in Chem C (Table A7), Pearson correlations using two different measures of language comprehension produced very similar results (r = +.40 - +.45). As expected, in Chem C, the two different measures of comprehension ability, SAT-CR section scores and GMRT scores, were very strongly correlated (r = +.65). In all cases, comprehension ability (regardless of the measure used) was strongly correlated with SAT-Math section scores (r = +.41 - +.58). Correlations of comprehension ability and prior knowledge were also statistically significant with medium to large effect sizes (r = +.36 - +.53). These correlations however did not indicate extremely high multicollinearity among comprehension ability measures and other predictor variables used in this study.

Correlational analyses were also performed in order to establish relationships between course performance, math ability, and prior chemistry knowledge. Both math ability and prior knowledge had positive correlations with exam performance that corresponded to mostly medium effect sizes in each of the courses. Only one correlation (that of prior knowledge and ACS exam score in Chem B, Table A6) differed and would be deemed “large”. This result was consistent with Chem B being the second course of a two-semester general chemistry sequence.

Table A5. Correlations of performance predictors and ACS exam for Chem A.

	SAT-Critical Reading	*ACS Toledo Placement Exam*	*ACS First Term General Chemistry Paired Questions Exam*
SAT-Math	.550* (N = 1124)	.493* (N = 1054)	.447* (N = 805)
ACS Toledo Placement Exam	.374* (N = 1054)	-----	.386* (N = 872)
ACS First Term General Chemistry Paired Questions Exam	.349* (N = 806)	-----	-----

*$p < .001$

Table A6. Correlations of performance predictors and ACS exam for Chem B.

	SAT-Math	*SAT-Critical Reading*	*ACS General Chemistry Conceptual Exam (Part I)*	*ACS Special Exam (1997)*
SAT-Math	-----	.579* (N = 497)	.560* (N = 462)	.433* (N = 489)
SAT-Critical Reading	-----	-----	.528* (N = 462)	.454* (N = 489)
ACS General Chemistry Conceptual Exam (Part I)	-----	-----	-----	.528* (N = 510)

*$p < .001$

Table A7. Correlations of performance predictors and ACS exam for Chem C.

	SAT-Critical Reading	*SAT-Math*	*ACS Toledo Placement Exam*	*ACS General Chemistry (Conceptual) Exam*
Gates-MacGinitie Reading Test	.645* (*N* = 488)	.411* (*N* = 489)	.382* (*N* = 534)	.396* (*N* = 528)
SAT-Math	.501* (*N* = 536)	-----	.439* (*N* = 503)	.475* (*N* = 528)
ACS Toledo Placement Exam	.363* (*N* = 502)	-----	-----	.447* (*N* = 541)
ACS General Chemistry (Conceptual) Exam	.451* (*N* = 528)	-----	-----	-----

* $p < .001$

Appendix 3. Evaluation of intraclass correlations for Chem A and Chem B.

In Chem A, variance parameter estimates indicated that between-classroom effects accounted for relatively little variance in ACS exam score as compared to between-students variance (s^2_{bg} = 2.51, s^2_{wg} = 260.43, ρ = 0.010). We concluded from this that individual classrooms had little effect on the relationships investigated in this study and that analysis of these data by hierarchical linear modeling was not necessary. Thus, linear regression techniques were chosen to analyze ACS exam data in course A.

In Chem B, the intraclass correlation was actually quite large (s^2_{bg} = 110.15, s^2_{wg} = 244.23, ρ = 0.311). However, the standard error of the between-groups variance was also very large (91.12), leading to statistical insignificance of the correlation. We thus concluded from this that individual classrooms had little effect on the relationships investigated in Chem B and that analysis of these data by hierarchical linear modeling was not necessary. Thus, linear regression techniques were chosen to analyze ACS exam data in course B.

Appendix 4. Hierarchical linear modeling of Course C midterm exam data to address Research Question 1.

As we collected four exam scores for each student in Chem C, we cannot assume that any of these observations is independent of the other. We must assume that each exam score is dependent on those that came before it, i.e. related by time. Thus, a 2-level linear growth model was used to analyze these midterm exam data (Raudenbush & Bryk, 2002, pp 163-169). Because time was the only within-student variable considered in this study, this variable was entered as the only direct predictor of midterm exam score at level-1. The following regression equation described level-1:

$$\mathrm{MES}_{ij} = \pi_{0j} + \pi_{1j}(\mathrm{time}) + e_{ij} \tag{1}$$

where MES_{ij} was the standardized score on midterm exam i for student j, "time" was an ordinal variable that spanned from 0 (representing the first exam) to 3 (representing the fourth exam), and π_{0j}, the intercept, was the initial status (i.e. first exam score) for student j. The slope representing the mean rate of change in MES over time is given by π_{1j}. Deviations of individual exam scores for student j from π_{0j} are represented by e_{ij}.

Level-2 equations described the effect of a student-level parameter (comprehension ability) on performance and were constructed to predict the intercept and slope of Equation 1 from standardized language comprehension ability (LC, either SAT-CR or GMRT) scores:

$$\begin{aligned} \pi_{0j} &= \beta_{00} + \beta_{01}(\mathrm{LC}) + r_{0j} \\ \pi_{1j} &= \beta_{10} + \beta_{11}(\mathrm{LC}) + r_{1j} \end{aligned} \tag{2}$$

The level-2 intercept is represented by β_{00}, the mean initial status across all students. The mean growth rate is given by β_{10}. The slope representing the mean effect of language comprehension ability on exam score is given by β_{01}. The slope representing the effect of language comprehension on growth rate is given by β_{11}. Deviation of student j's mean exam score from the overall grand mean is given by r_{0j}. Finally, deviation of student j's growth rate from the overall mean growth rate is given by r_{1j}.

These models converged using the maximum likelihood (ML) estimation algorithm and residuals for all models followed a normal distribution, with means of approximately zero and standard deviations of ~0.5.

In order to evaluate the overall fit of the models, general linear hypothesis testing using the –2 Log Likelihood (–2LL) statistic was performed (Tabachnik & Fidell, 2013, pp 834-835). This can be done because the difference in the –2LL statistics of two models follows a chi-squared distribution, with degrees of freedom equal to the difference in the number of parameters between the two models. Full models (presented in the main text) were significantly better than unconditional growth models (i.e. ones in which only the intercepts and time were included). The results of these chi-squared tests are presented in Tables A8 and A9. For the model using SAT-CR scores as the measure of comprehension ability, χ^2 (2, N = 2130) = 4992.696 – 4895.532 = 97.164, $p < .001$. For the model using GMRT scores as the measure of comprehension ability, χ^2 (2, N = 2131) = 4955.924 – 4955.924 = 92.248, $p < .001$.

Table A8. Comparison of hierarchical linear models for midterm exam scores over time due to language comprehension ability (measured by SAT-CR scores).

Model	–2 Log Likelihood	df	χ^2 Difference Test
Unconditional growth	4992.696	6	
Final	4895.532	8	M2 – M1 = 97.164*

* $p < 0.001$

Table A9. Comparison of hierarchical linear models for midterm exam scores over time due to language comprehension ability (measured by GMRT scores).

Model	–2 Log Likelihood	df	χ^2 Difference Test
Unconditional growth	4955.924	6	
Final	4863.676	8	M2 – M1 = 92.248*

* $p < 0.001$

Appendix 5. Summary of multiple regression models of comprehension ability (as measured by SAT-CR section scores) and math ability as predictors of course performance (as measured by ACS exam scores) in Chem A and Chem B.

Regression analyses were employed to examine how comprehension ability and math ability compared as predictors of performance in Chem A and Chem B. The regression models were also used to account for variance uniquely predicted by comprehension ability and math ability as well as variance shared by these parameters. In both courses, these models explained a significant portion of variance in ACS exam score. As in Chem C (discussed in the main manuscript), both math ability and comprehension ability predicted significant increases in ACS exam score when the other was statistically controlled. In Chem A (Table A10), math ability had the larger effect on ACS exam score; however, the opposite was true in Chem B (Table A11). We found this result in Chem B to be interesting, given that much of the content of Chem B (thermodynamics, kinetics, chemical equilibrium, etc.) is assumed to be more demanding of students knowledge of math than the content of Chem A.

Since the squared semipartials (sr^2) for comprehension ability and math ability represent the unique variance of that predictor shared with ACS exam score, the sum of the squared semipartials can be subtracted from the overall R^2 for the regression model to determine the amount of variance common to both predictors. Table A10 presents a regression model using SAT-Math and SAT-CR scores as predictors of ACS Exam score in Chem A that accounted for 21.3% of the total variance in the outcome measure. SAT-Math scores uniquely predicted 9.4% of the total variance explained by the model, while SAT-CR scores uniquely predicted 1.3% of the total variance. The 10.7% of variance in ACS exam score remaining (roughly half of the total) must be shared equally by comprehension ability and math ability. Table A11 presents a regression model using SAT-Math and SAT-CR scores as predictors of ACS exam score in Chem B that accounted for 25.1% of the total variance in the outcome measure. SAT-Math scores uniquely predicted 4.5% of the total variance explained by the model, while SAT-CR scores uniquely predicted 6.4% of the total variance. The 14.2% of variance in ACS exam score remaining (again, roughly half of the total) must be shared equally by comprehension ability and math ability. In both Chem A and Chem B, the proportion of variance in ACS exam score shared by math ability and comprehension ability was consistently larger than the variance uniquely

predicted by either variable. This redundancy of comprehension ability and prior knowledge is very similar to that observed in Chem C, discussed in the main manuscript

Table A10. Standard multiple regression of SAT-Math scores and SAT-CR scores on Chem A course performance (as measured by the ACS First Term General Chemistry Paired Questions Exam). $N = 804$.

coefficient	*B*	*standard error*	β	*t*	sr^2	*p*
intercept	63.42	.507		125.091		< .001
SAT-CR	.029	.008	.138	3.641	.013	< .001
SAT-Math	.083	.009	.370	9.759	.094	< .001

$R^2 = .213$, $F(2,801) = 108.276$, $p < .001$

Table A11. Standard multiple regression of SAT-Math scores and SAT-CR scores on Chem B course performance (as measured by the ACS Special Exam (1997)). $N = 489$.

coefficient	*B*	*standard error*	β	*t*	sr^2	*p*
intercept	64.59	.727		88.804		< .001
SAT-CR	.072	.011	.306	6.408	.064	< .001
SAT-Math	.065	.012	.258	5.403	.045	< .001

$R^2 = .251$, $F(2,486) = 81.295$, $p < .001$

Appendix 6. Hierarchical linear modeling of instructor-generated midterm exam data to address Research Question 2.

As was done to address Research Question 1 (Appendix 3), a 2-level linear growth model was used to analyze midterm exam data. Because time was the only within-student variable considered in our study, it was entered as the only direct predictor of midterm exam score at level-1. Level-2 equations were then constructed to predict the intercept of the level-1 from standardized language comprehension scores (LC, either SAT-CR or GMRT scores) as well as standardized SAT-Math scores (SAT-M):

$$\begin{aligned}
&\underline{\text{Level-1}} \\
&\qquad \text{MES}_{ij} = \pi_{0j} + \pi_{1j}(\text{time}) + e_{ij} \\
&\underline{\text{Level-2}} \\
&\qquad \pi_{0j} = \beta_{00} + \beta_{01}(\text{LC}) + \beta_{02}(\text{SAT-M}) + r_{0j} \\
&\qquad \pi_{1j} = \beta_{10} + \beta_{11}(\text{LC}) + \beta_{12}(\text{SAT-M}) + r_{1j}
\end{aligned} \tag{3}$$

The level-2 intercept is represented by β_{00}, the mean initial status across all students. The mean growth rate is given by β_{10}. The slope representing the mean effect of language comprehension ability on exam score is given by β_{01}, while the slope representing the mean effect of math ability is given by β_{02}. The slope representing the effect of language comprehension on growth rate is given by β_{11} and the slope representing the effect of math ability on growth rate is given by β_{12}. Deviation of student *j*'s mean exam score from the overall mean initial status is given by r_{0j}. Finally, deviation of student *j*'s growth rate from the overall mean growth rate is given by r_{1j}.

These models converged using the maximum likelihood (ML) estimation algorithm and residuals for both models followed a normal distribution, with means of approximately zero and standard deviations of ~0.5. Results for the full model using SAT-CR scores as the measure of comprehension ability are presented in Tables A12-A13, while those for the model using GMRT scores are presented in Tables A14 – A15.

Table A12. Estimation of the effects of language comprehension (as measured by SAT-CR section scores) and math ability on Chem C course performance (as measured by course midterm exams).

Fixed Effect	*estimate*	*standard error*	*approx. df*	*t*	*p*
mean initial status, β_{00}	.021	.035	533	.599	.549
mean growth rate, β_{10}	–.006	.012	531	–.507	.612
SAT-CR score effect, β_{01}	.191	.041	533	4.686	< 0.001
SAT-Math score effect, β_{02}	.350	.041	535	8.556	< 0.001
SAT-CR score by time, β_{11}	–.011	.014	529	–.770	.441
SAT-Math score by time, β_{12}	–.022	.014	531	–1.562	.119
Random Effect	*variance*	*standard error*	*Wald Z*	*p*	
level-1: exam scores (residual, e_{ij})	.348	.015	22.928	<0.001	
level-2: student (initial status, r_{0j})	.405	.041	9.825	<0.001	
level-2: student (covariance)	.003	.012	.252	.801	
level-2: student (growth rate, r_{1j})	.009	.006	1.615	.106	

This model was significantly better than an unconditional growth model—the results of the chi-squared test describing this are presented in Table A13. For the model using SAT-CR scores as the measure of comprehension ability, χ^2 (4, N = 2126) = 4976.022 – 4806.855 = 169.167, p < .001.

Table A13. Comparison of hierarchical linear models for midterm exam scores over time due to language comprehension ability (measured by SAT-CR scores) and math ability.

Model	–2 Log Likelihood	df	χ^2 Difference Test
Unconditional growth	4976.022	6	
Final	4806.855	10	M2 – M1 = 169.167*

* p < 0.001

Table A14. Estimation of the effects of language comprehension (as measured by GMRT scores) and math ability on Chem C course performance (as measured by course midterm exams).

Fixed Effect	*estimate*	*standard error*	*approx. df*	*t*	*p*
mean initial status, β_{00}	.022	.036	490	.615	.539
mean growth rate, β_{10}	–.003	.013	486	–.242	.809
GMRT score effect, β_{01}	.211	.041	490	5.176	< 0.001
SAT-Math score effect, β_{02}	.361	.041	492	8.850	< 0.001
GMRT score by time, β_{11}	–.029	.014	487	–2.043	.042
SAT-Math score by time, β_{12}	–.013	.014	489	–.943	.346
Random Effect	*variance*	*standard error*	*Wald Z*	*p*	
level-1: exam scores (residual, e_{ij})	.344	.016	21.972	<0.001	
level-2: student (initial status, r_{0j})	.407	.043	9.499	<0.001	
level-2: student (covariance)	.005	.012	.425	.671	
level-2: student (growth rate, r_{1j})	.009	.006	1.510	.131	

This model was significantly better than an unconditional growth model—the results of the chi-squared test describing this are presented in Table A15. For the model using GMRT scores as the measure of comprehension ability, χ^2 (4, N = 1947) = 4545.979 – 4387.674 = 158.305, p < .001.

Table A15. Comparison of hierarchical linear models for midterm exam scores over time due to language comprehension ability (measured by GMRT scores) and math ability.

Model	–2 Log Likelihood	df	χ^2 Difference Test
Unconditional growth	4545.979	6	
Final	4387.674	10	M2 – M1 = 158.305*

* p < 0.001

Appendix 7. Regression model summaries with the prior knowledge and SAT-Critical Reading scores as predictors of performance in Chem A and Chem B.

Research Question 3 posited that high comprehension ability compensates for low prior knowledge in general chemistry courses. To help address this question, sequential regressions were employed to determine if the addition of comprehension ability by prior knowledge interaction terms improved predictions of ACS exam score beyond that afforded by differences in comprehension ability and prior knowledge. Tables A16 and A17 display the final results of these analyses.

For Chem A, after step-1, with SAT-CR and ACS Toledo exam scores entered in the regression, R^2 = .193, $F_{inc}(2,736)$ = 88.163, p <.001. After step-2, with the SAT-CR by Toledo interaction added to the prediction of ACS exam score, R^2 = .194, $F_{inc}(1,735)$ = .358, p =.550. Therefore, addition of the SAT-CR by Toledo interaction to the regression with SAT-CR and Toledo Exam scores did not result in a significant increase in R^2. This pattern of results suggested that SAT-CR and Toledo exam scores predicted approximately a fifth of the variability in ACS exam score; an interaction between these two variables did not contribute to that prediction.

For the case of Chem B, after step-1, with SAT-CR and ACS General Chemistry Conceptual Exam (Part I) (GCC) scores entered in the regression, R^2 = .332, $F_{inc}(2,452)$ = 112.168, p <.001. After step-2, with the SAT-CR by GCC interaction added to the prediction of ACS exam score, R^2 = .333, $F_{inc}(1,451)$ = 1.188, p =.276. Therefore, addition of the SAT-CR by GCC exam interaction to the regression with SAT-CR and GCC Exam scores did not result in a significant increase in R^2. This pattern of results suggested that SAT-CR and GCC exam scores predicted approximately a third of the variability in ACS Exam score; an interaction between these two variables did not contribute to that prediction.

Thus, in both Chem A and Chem B (as in Chem C, discussed in the main text), students with high prior knowledge and high comprehension ability scored better on ACS exams than students with low prior knowledge and the same level of high comprehension ability. However, given the significant main effect of comprehension ability—students with low prior knowledge and high comprehension ability scored better on ACS exams than students with low prior knowledge and low comprehension ability. This is illustrated in Figure A1, which plots standardized ACS exam score versus comprehension ability for students of low- and high-prior

knowledge in both Chem A and Chem B. Figure A1 was constructed using the regression coefficients found in Tables A16 and A17. The large performance gap between students of low and high levels of prior chemistry knowledge is clearly illustrated in this Figure, as is the ability of comprehension ability to potentially compensate for low prior knowledge. For example, in Chem A, students possessing low prior knowledge and high comprehension ability scored approximately the same as students possessing high prior knowledge and low comprehension ability. Thus, high comprehension ability can help students negotiate the achievement gap present between those possessing low and high levels of prior chemistry knowledge. The compensatory ability of language comprehension is somewhat more tenuous in Chem B, where the performance gap between students possessing low and high levels of prior knowledge is much wider.

Table A16. Sequential regression of SAT-CR and ACS Toledo Exam scores on Chem A course performance (as measured by the ACS First Term General Chemistry Paired Questions Exam). $N = 739$.

coefficient	*B*	*stand. error*	β	*t*	sr^2	*p*
intercept	63.708	.572		111.441		< .001
SAT-CR (step-1)	.048	.008	.228	6.371	.044	< .001
Toledo Exam (step-1)	.405	.049	.302	8.285	.075	< .001
SAT-CR by Toledo (step-2)	<.001	.001	–.020	–.598	.0004	.550
$R^2 = .194$, $F(3,735) = 58.843$, $p < .001$						

Table A17. Sequential regression of SAT-CR and ACS General Chemistry Conceptual Exam (Part I) scores on Chem B course performance (as measured by the ACS Special Exam (1997)). $N = 455$.

coefficient	*B*	*stand. error*	*β*	*t*	sr^2	*p*
intercept	63.87	.801		79.742		< .001
SAT-CR (step-1)	.054	.001	.228	5.001	.037	< .001
ACS General Chemistry Conceptual Exam (Part I) (step-1)	.480	.055	.405	8.679	.112	< .001
SAT-CR by GGC Exam (step-2)	.001	.001	.045	1.090	.002	.276
$R^2 = .333$, $F(3,451) = 75.206$, $p < .001$						

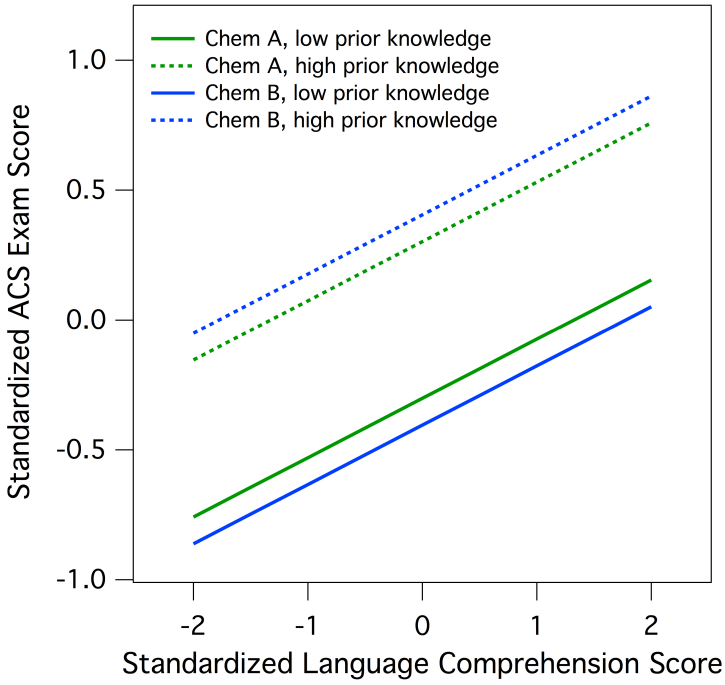

Figure A1. A plot of language comprehension ability vs. predicted ACS exam score for students of low (solid) and high (dotted) prior knowledge in Chem A (green) and Chem B (blue). This plot was constructed using the standardized regression coefficients listed in Tables A16 and A17. Low prior knowledge students were modeled as achieving scores one standard deviation below the mean on prior knowledge measures, while high prior knowledge students were modeled as achieving one standard deviation above the mean on prior knowledge measures.

Appendix 8. Hierarchical linear modeling of instructor-generated midterm exam data to address Research Question 3.

As for the previous Research Questions (Appendices 3 and 5), a 2-level linear growth model was used to analyze midterm exam data. Because time was the only within-student variable considered in our study, it was entered as the only direct predictor of midterm exam score at level-1. Level-2 equations were then constructed to predict the intercept of the level-1 from standardized language comprehension scores (LC, either SAT-CR or GMRT scores), standardized Toledo Exam scores (Toledo), and a language comprehension by prior knowledge interaction term (LC × Toledo):

$$\begin{aligned}
&\underline{\text{Level-1}} \\
&\quad \text{MES}_{ij} = \pi_{0j} + \pi_{1j}(\text{time}) + e_{ij} \\
&\underline{\text{Level-2}} \\
&\quad \pi_{0j} = \beta_{00} + \beta_{01}(\text{LC}) + \beta_{02}(\text{Toledo}) + \beta_{03}(\text{LC} \times \text{Toledo}) + r_{0j} \\
&\quad \pi_{1j} = \beta_{10} + \beta_{11}(\text{LC}) + \beta_{12}(\text{Toledo}) + \beta_{13}(\text{LC} \times \text{Toledo}) + r_{1j}
\end{aligned} \tag{4}$$

The level-2 intercept is represented by β_{00}, the mean initial status across all students. The mean growth rate is given by β_{10}. The slope representing the mean effect of language comprehension ability on exam score is given by β_{01}; the slope representing the mean effect of prior knowledge is given by β_{02}; and the slope representing the mean effect of language comprehension by prior knowledge moderation is given by β_{03}. The slope representing the effect of language comprehension on growth rate is given by β_{11}; the slope representing the effect of prior knowledge on growth rate is given by β_{12}; and the slope representing the effect of language comprehension by prior knowledge moderation on growth rate is given by β_{13}. Deviation of student *j*'s mean exam score from the overall grand mean is given by r_{0j}. Finally, deviation of student *j*'s growth rate from the overall mean growth rate is given by r_{1j}.

These models converged using the maximum likelihood (ML) estimation algorithm and residuals for both models followed a normal distribution, with means of approximately zero and standard deviations of ~0.5. Results for the full model using SAT-CR scores as the measure of comprehension ability are presented in Tables A18-A19, while those for the model using GMRT scores are presented in Tables A20 – A21.

Table A18. Full model of the effects of language comprehension (as measured by SAT-CR section scores) and prior knowledge on Chem C course performance (as measured by course midterm exams), including hypothesized interaction terms.

Fixed Effect	*estimate*	*standard error*	*approx. df*	*t*	*p*
mean initial status, β_{00}	.023	.037	502	.694	.488
mean growth rate, β_{10}	–.004	.013	499	–.284	.777
SAT-CR score effect, β_{01}	.225	.039	503	5.810	< 0.001
Toledo Exam score effect, β_{02}	.369	.039	503	9.409	< 0.001
SAT-CR score by time, β_{11}	–.013	.014	500	–.950	.342
Toledo Exam score by time, β_{12}	–.025	.014	496	–1.793	.074
SAT-CR score by Toledo score, β_{03}	–.003	.035	506	–.080	.936
interaction by time, β_{13}	.0005	.012	502	.037	.971
Random Effect	*variance*	*standard error*	*Wald Z*	*p*	
level-1: exam scores (residual, e_{ij})	.344	.015	22.246	<0.001	
level-2: student (initial status, r_{0j})	.388	.041	9.423	<0.001	
level-2: student (covariance)	.004	.012	.368	.713	
level-2: student (growth rate, r_{1j})	.008	.006	1.447	.148	

This model was significantly better than an unconditional growth model, χ^2 (6, N = 1934) = 4649.673 – 4477.132 = 172.541, p < .001. However, coefficients relating to the hypothesized language comprehension by prior knowledge interaction were clearly not significant. Therefore, a final growth model that did not include this interaction was proposed. This final model did not differ significantly from the full model, χ^2 (2, N = 1934) = 4477.138 – 4477.132 = .006, p > .05; in fact, the coefficients, standard errors, and associated statistics for the final model did not differ from those reported in Table A18. Table A19 summarizes the models evaluated.

Table A19. Comparison of hierarchical linear models for midterm exam scores over time due to language comprehension ability (measured by SAT-CR scores) and prior knowledge.

Model	–2 Log Likelihood	df	χ^2 Difference Test
Unconditional growth	4649.673	6	
Full (including SAT-CR × Toledo)	4477.132	12	M2 – M1 = 172.541*
Final	4477.138	10	M3 – M2 = .006

* $p < 0.001$

Table A20. Full model of the effects of language comprehension (as measured by GMRT section scores) and prior knowledge on Chem C course performance (as measured by course midterm exams), including hypothesized interaction terms.

Fixed Effect	*estimate*	*standard error*	*approx. df*	*t*	*p*
mean initial status, β_{00}	.022	.037	533	.603	.547
mean growth rate, β_{10}	–.002	.013	528	–.163	.871
GMRT score effect, β_{01}	.237	.038	534	6.265	< 0.001
Toledo Exam score effect, β_{02}	.344	.037	534	9.226	< 0.001
GMRT score by time, β_{11}	–.031	.013	532	–2.365	.018
Toledo Exam score by time, β_{12}	–.019	.013	526	–1.441	.150
GMRT score by Toledo score, β_{03}	–.022	.035	538	–.642	.521
interaction by time, β_{13}	.0083	.012	539	.667	.505
Random Effect	*variance*	*standard error*	*Wald Z*	*p*	
level-1: exam scores (residual, e_{ij})	.340	.015	22.872	<0.001	
level-2: student (initial status, r_{0j})	.398	.040	9.860	<0.001	
level-2: student (covariance)	.006	.012	.521	.602	
level-2: student (growth rate, r_{1j})	.009	.006	1.535	.125	

This model was significantly better than an unconditional growth model, χ^2 (6, N = 2115) = 4918.454 – 4741.393 = 177.061, p < .001. However, coefficients relating to the hypothesized language comprehension by prior knowledge interaction were clearly not significant. Therefore, a final growth model that did not include this interaction was proposed. This final model did not differ significantly from the full model, χ^2 (2, N = 2115) = 4741.991 – 4741.393 = .598, p > .05; in fact, the coefficients, standard errors, and associated statistics for the final model did not differ from those reported in Table A20. Table A21 summarizes the models evaluated.

Table A21. Comparison of hierarchical linear models for midterm exam scores over time due to language comprehension ability (measured by GMRT scores) and prior knowledge.

Model	–2 Log Likelihood	df	χ^2 Difference Test
Unconditional growth	4918.454	6	
Full (including GMRT × Toledo)	4741.393	12	M2 – M1 = 177.061*
Final	4741.991	10	M3 – M2 = .598

* p < 0.001

Appendix 9. Summary of multiple regression models of comprehension ability, math ability, and prior knowledge as predictors of course performance (as measured by ACS exam scores) in Chem A, Chem B, and Chem C.

Regression analyses were employed to examine models predicting achievement in Chem A, Chem B, and Chem C from comprehension ability, prior knowledge, and math ability. Thus, course performance (as measured by ACS exams) was regressed on mean-centered versions of these variables (and on a comprehension ability by prior knowledge interaction term, as hypothesized in earlier models) in all three courses. Regardless of chemistry course or (in the case of Chem C), measure of comprehension ability, these models explained a significant portion of variance in ACS exam score (R^2 = .24 - .34). Also in all cases, comprehension ability, prior knowledge, and math ability predicted significant increases in ACS exam score when the others were statistically controlled. In Chem A (Table A22) and Chem C (Tables A24 and A25), math ability had the largest effect on ACS exam score, while comprehension ability had the smallest. In Chem B (Table A23), prior knowledge had the largest effect on ACS exam score, while math ability had the smallest. This is consistent with Chem B being the second course in a two-semester general chemistry sequence. In only one case (Chem C in which comprehension ability was measured with SAT-CR scores) was the prior knowledge by comprehension ability interaction term statistically significant. However, similar to the other instance in which this interaction tested significant (Table 9, main text), its effect was too small to be considered meaningful.

Table A22. Standard multiple regression of SAT-Math scores, SAT-CR scores, and Toledo Exam scores on Chem A course performance (as measured by the ACS First Term General Chemistry Paired Questions Exam). $N = 737$.

coefficient	*B*	*standard error*	β	*t*	sr^2	*p*
intercept	63.496	.557		114.067		< .001
SAT-CR Score	.023	.008	.108	2.766	.008	.006
Toledo Exam Score	.278	.051	.207	5.443	.031	< .001
SAT-Math Score	.064	.009	.281	5.786	.048	< .001
SAT-CR by Toledo	<.001	.001	–.017	–.526	.0003	.599
$R^2 = .238$, $F(4,732) = 57.091$, $p < .001$						

Table A23. Standard multiple regression of SAT-Math scores, SAT-CR scores, and ACS General Chemistry Conceptual Exam (Part I) scores on Chem B course performance (as measured by the ACS Special Exam (1997)). $N = 455$.

coefficient	*B*	*standard error*	β	*t*	sr^2	*p*
intercept	63.85	.796		80.22		< .001
SAT-CR Score	.043	.012	.180	3.664	.020	< .001
GCC Exam Score	.424	.059	.358	7.186	.076	< .001
SAT-Math Score	.033	.013	.129	2.598	.010	.010
SAT-CR by GCC	.001	.001	.047	1.152	.002	.250
$R^2 = .343$, $F(4,450) = 58.811$, $p < .001$						

Table A24. Standard multiple regression of SAT-Math scores, SAT-CR scores, and Toledo Exam scores on Chem C course performance (as measured by the ACS General Chemistry (Conceptual) Exam). N = 497.

coefficient	*B*	*standard error*	β	*t*	sr^2	*p*
intercept	64.191	.486		125.957		< .001
SAT-CR Score	.043	.008	.239	5.538	.042	< .001
Toledo Exam Score	.221	.044	.211	5.058	.035	< .001
SAT-Math Score	.053	.009	.265	5.927	.048	< .001
SAT-CR by Toledo	.001	.001	.085	2.264	.007	.024
R^2 = .333, F(4,492) = 61.358, p < .001						

Table A25. Standard multiple regression of SAT-Math scores, GMRT scores, and Toledo Exam scores on Chem C course performance (as measured by the ACS General Chemistry (Conceptual) Exam). N = 482.

coefficient	*B*	*standard error*	β	*t*	sr^2	*p*
intercept	64.191	.486		125.957		< .001
GMRT Score	.142	.037	.167	3.871	.022	< .001
Toledo Exam Score	.222	.046	.210	4.796	.033	< .001
SAT-Math	.065	.009	.321	7.234	.076	< .001
GMRT by Toledo	.004	.003	.052	1.637	.003	.172
R^2 = .311, F(4,447) = 53.757, p < .001						

Appendix 10. Evaluation of models predicting course performance (as measured by instructor-generated midterm exams in Chem C) from comprehension ability, prior knowledge, and math ability using hierarchical linear modeling.

As for previous models of Chem C midterm exam data (Appendices 3, 5, and 8), a 2-level linear growth model was used. Because time was the only within-student variable considered in our study, it was entered as the only direct predictor of midterm exam score at level-1. Level-2 equations were then constructed to predict the intercept of the level-1 from standardized language comprehension scores (LC, either SAT-CR or GMRT scores), standardized Toledo Exam scores (Toledo), standardized SAT-Math scores, and a language comprehension by prior knowledge interaction term (LC × Toledo):

Level-1

$$\text{MES}_{ij} = \pi_{0j} + \pi_{1j}(\text{time}) + e_{ij} \tag{5}$$

Level-2

$$\pi_{0j} = \beta_{00} + \beta_{01}(\text{LC}) + \beta_{02}(\text{Toledo}) + \beta_{03}(\text{SAT-Math}) + \beta_{04}(\text{LC} \times \text{Toledo}) + r_{0j}$$

$$\pi_{1j} = \beta_{10} + \beta_{11}(\text{LC}) + \beta_{12}(\text{Toledo}) + \beta_{13}(\text{SAT-Math}) + \beta_{14}(\text{LC} \times \text{Toledo}) + r_{1j}$$

The level-2 intercept is represented by β_{00}, the mean initial status across all students. The mean growth rate is given by β_{10}. The slope representing the mean effect of language comprehension ability on exam score is given by β_{01}; the slope representing the mean effect of prior knowledge is given by β_{02}; the slope representing the mean effect of math ability is given by $\beta_{03;}$ and the slope representing the mean effect of language comprehension by prior knowledge moderation is given by β_{04}. The slope representing the effect of language comprehension on growth rate is given by β_{11}; the slope representing the effect of prior knowledge on growth rate is given by β_{12}; the slope representing the mean effect of math ability on growth rate is given by β_{13}; and the slope representing the effect of language comprehension by prior knowledge moderation on growth rate is given by β_{14}. Deviation of student *j*'s mean exam score from the overall grand mean is given by r_{0j}. Finally, deviation of student *j*'s growth rate from the overall mean growth rate is given by r_{1j}.

These models converged using the maximum likelihood (ML) estimation algorithm and residuals for both models followed a normal distribution, with means of approximately zero and

standard deviations of ~0.5. Results for the full model using SAT-CR scores as the measure of comprehension ability are presented in Tables A26 and A27, while those for the model using GMRT scores are presented in Tables A28 and A29.

Table A26. Full model of the effects of language comprehension (as measured by SAT-CR section scores), prior knowledge, and math ability on Chem C course performance (as measured by course midterm exams), including hypothesized interaction terms.

Fixed Effect	*estimate*	*standard error*	*approx. df*	*t*	*p*
mean initial status, β_{00}	.020	.036	503	.566	.572
mean growth rate, β_{10}	–.004	.013	499	–.267	.790
SAT-CR score effect, β_{01}	.126	.041	503	3.102	.002
Toledo Exam score effect, β_{02}	.288	.040	502	7.198	< .001
SAT-Math score effect, β_{03}	.264	.042	505	6.206	< .001
SAT-CR score by Toledo score, β_{04}	.012	.034	506	.337	.737
SAT-CR score by time, β_{11}	–.007	.015	498	–.470	.639
Toledo Exam score by time, β_{12}	–.020	.015	496	–1.368	.172
SAT-Math score by time, β_{13}	–.016	.016	500	–1.025	.306
interaction by time, β_{14}	–.0007	.012	502	–.053	.957
Random Effect	*variance*	*standard error*	*Wald Z*	*p*	
level-1: exam scores (residual, e_{ij})	.344	.015	22.245	<0.001	
level-2: student (initial status, r_{0j})	.343	.038	8.923	<0.001	
level-2: student (covariance)	.007	.011	.623	.534	
level-2: student (growth rate, r_{1j})	.008	.006	1.414	.157	

The model using SAT-CR scores as the measure of comprehension ability was significantly better than an unconditional growth model, χ^2 (8, N = 1996) = 4649.673 – 4436.277 = 213.396, p < .001. However, coefficients relating to the hypothesized language comprehension

by prior knowledge interaction were clearly not significant. Therefore, a final growth model that did not include this interaction was proposed. This final model did not differ significantly from the full model, χ^2 (2, N = 1996) = 4436.277 – 4436.403 = .126, $p > .05$; in fact, the coefficients, standard errors, and associated statistics for the final model did not differ substantially from those reported in Table A26. Table A27 summarizes the models evaluated.

Table A27. Comparison of hierarchical linear models for midterm exam scores over time due to language comprehension ability (measured by SAT-CR scores), prior knowledge, and math ability.

Model	–2 Log Likelihood	df	χ^2 Difference Test
Unconditional growth	4649.673	6	
Full (including SAT-CR × Toledo)	4436.277	14	M2 – M1 = 213.396*
Final	4436.403	12	M3 – M2 = .126

* $p < 0.001$

Similar to the regression model using SAT-CR scores to predict ACS exam score in Chem C (Table A24), comprehension ability, prior knowledge, and math ability predicted significant increases in midterm exam score when the others were statistically controlled. Unlike the associated regression model, however, prior knowledge had the largest effect on midterm exam score; the smallest effect remained associated with comprehension ability. No significant effects on growth rate were observed for any of the predictors.

Table A28. Full model of the effects of language comprehension (as measured by GMRT section scores), prior knowledge, and math ability on Chem C course performance (as measured by course midterm exams), including hypothesized interaction terms.

Fixed Effect	*estimate*	*standard error*	*approx. df*	*t*	*p*
mean initial status, β_{00}	.028	.037	487	.756	.450
mean growth rate, β_{10}	–.0009	.013	483	–.064	.949
GMRT score effect, β_{01}	.144	.040	487	3.586	< .001
Toledo Exam score effect, β_{02}	.282	.041	486	6.862	< .001
SAT-Math score effect, β_{03}	.260	.042	488	6.231	< .001
GMRT score by Toledo score, β_{04}	–.020	.037	490	–.546	.586
GMRT score by time, β_{11}	–.029	.015	485	–1.980	.048
Toledo Exam score by time, β_{12}	–.012	.015	480	–.799	.425
SAT-Math score by time, β_{13}	–.009	.105	485	–.569	.569
interaction by time, β_{14}	–.002	.013	485	–.114	.909
Random Effect	*variance*	*standard error*	*Wald Z*	*p*	
level-1: exam scores (residual, e_{ij})	.344	.016	21.905	<0.001	
level-2: student (initial status, r_{0j})	.345	.039	8.811	<0.001	
level-2: student (covariance)	.009	.012	.750	.453	
level-2: student (growth rate, r_{1j})	.008	.006	1.397	1.62	

The model using GMRT scores as the measure of comprehension ability was significantly better than an unconditional growth model, χ^2 (8, N = 1935) = 4517.729 – 4307.367 = 207.362, p < .001. However, coefficients relating to the hypothesized language comprehension by prior knowledge interaction were clearly not significant. Therefore, a final growth model that did not include this interaction was proposed. This model did not differ significantly from the full model, χ^2 (2, N = 1935) = 4307.827 – 4307.367 = .460, p > .05; in fact, the coefficients,

standard errors, and associated statistics for the final model did not differ substantially from those reported in Table A28. Table A29 summarizes the models evaluated.

Table A29. Comparison of hierarchical linear models for midterm exam scores over time due to language comprehension ability (measured by GMRT scores), prior knowledge, and math ability.

Model	–2 Log Likelihood	df	χ^2 Difference Test
Unconditional growth	4514.729	6	
Full (including GMRT × Toledo)	4307.367	14	M2 – M1 = 207.362*
Final	4307.827	12	M3 – M2 = .460

* $p < 0.001$

Similar to the regression model using GMRT scores to predict ACS exam score in Chem C (Table A25), comprehension ability, prior knowledge, and math ability predicted significant increases in midterm exam score when the others were statistically controlled. Unlike the associated regression model, however, prior knowledge had the largest effect on midterm exam score; the smallest effect remained associated with comprehension ability. This result was similar to the HLM using SAT-CR scores as the measure of comprehension ability (Table A26). A significant effect of GMRT score on growth rate was also observed in this model, as it had been in all other HLMs using GMRT scores as a measure of comprehension ability.