Electronic Supplementary Information (ESI)

Table of Contents

1) Further description of institutions

A brief summary of the five institutions where assessment data was collected is included below in Table 1.

Table 1 Description of the institutions where data collection took place.

Institution	
Code	**Description**
I1	Small suburban comprehensive school with undergraduate only chemistry
I2	Medium rural predominantly undergraduate institution
I3	Large urban research-intensive institution
I4	Medium rural mastering level Hispanic serving school
I5	Professional school with organic followed by one term of general chemistry

2) Instructions provided to students completing the IFAT

"If your first scratch unveils a star, you've gotten the answer correct and you should proceed to the next question. If your first scratch unveils a blank square, you have not chosen the correct answer and you should reread the question and select/scratch another answer. Repeat this process until you uncover the star representing the correct answer. Please only circle the FIRST scratched choice for each question not any of the 2nd, 3rd, or 4th scratches."

3) Expanded statistical background

3.1) Hierarchal Linear Modelling (HLM)

HLM's are commonly used in a variety of fields and it is in-part because of their diverse use that the nomenclature surrounding HLMs is often inconsistent (Singer, 1998; O'Connell and McCoach, 2004; Cornelius et al., 2007; Laursen and Weston, 2014). A few of the common names used to refer to HLM-type models are: multilevel models, nested data models, linear mixed-effect models, value-added models, etc. For the purposes of this study the models will only be referred to as HLMs and the models will be represented in a manner consistent with how it was formatted by Doran (Doran and Lockwood, 2006). One limitation with any type of linear model is the results will only be as accurate as the scores that are used to construct the model. With that in mind, to test validity, four models were constructed using different scoring techniques. One of these scoring techniques was the students true score which required item response theory (IRT) to estimate.

3.2) Item response theory

IRT has increased in popularity since the 20th century when it was first developed (Bock, 2005). Today, IRT is commonplace in psychometrics and is used in the development of major examinations such as the scholastic aptitude test (SAT) and graduate record examination (GRE) (An and Yung, 2014). The primary reason IRT has become such a cornerstone of psychometrics is because it uses student's responses to each of the items on the exam to estimate the students underlying ability (Cooper et al., 2008; Hambleton et al., 2012). An additional benefit of IRT is the prediction of students' abilities does not depend on the sample of students who took the exam which means IRT analysis will automatically account for any potential sampling error between treatments (Weaver and Sturtevant, 2015). However, construction of these IRT models comes at the cost of methodological simplicity and requires hefty sample sizes which for some research projects is not realistic (Glynn, 2012). For example, in this study sample sizes were not large enough to investigate specific treatment-level impacts with IRT, so the treatments needed to be grouped together for IRT results to be valid.

One possible expansion of IRT is the use of Lord's Wald test to investigate each question for the possibility of differential item functioning (DIF) (Lord, 1980). In the past, DIF has been primarily used in psychology and education for the purpose of investigating question bias between two groups (Kendhammer et al., 2013; Kendhammer and Murphy, 2014; Lee and Suh, 2018). While DIF analyses have typically been used to evaluate exam fairness for factors such as cultural or sex-based differences, these tests can equivalently be used to reveal when items perform differently before and after a treatment has been applied to the students (Holland and Wainer, 2009). After a test such as Lord's has been conducted, questions with a significant value only indicate that students perform differently on the exam before and after the treatment and post hoc analysis must be conducted to ensure that the treatment benefited the student (as opposed to harmed the students' performance). This post hoc analysis can be conducted in many different ways but use of item characteristic curves (ICCs) has the benefit of revealing differences at every student ability level (Zumbo, 1999).

3.3) Pilot HLM

To determine the optimal method for analyzing the data, four pilot models were constructed and are displayed in Table 2. These models are labeled m1-m4 and are in sequence based on increasing complexity. For interpretation of these models, Table 3 includes more details about the variables.

Table 2 Description of the variables used for all HLMs.

Symbol	Interpretation	Specific Symbol	Specific Interpretation
Y	Test score	Y_{ti}	Test score for student (i) at time (t)
β	Fixed effects	β_0	Average initial ability level of students during week 1
		β_1	Average student improvement from week 1 to week 2
		β_2	Semester main effect to account for variability among student-level intercepts
		β_3	Sex main effect to account for variability among student-level intercepts
θ	Treatment-Level Random Effects	$\theta_{0\,j(i)}$	Difference from average initial ability for a student (i) who underwent treatment (j)
		$\theta_{1\,j(i)}$	Difference from average student improvement for a student (i) who underwent treatment (j)
δ	Student-Level Random Effects	$\delta_{0\,i}$	Difference from average initial ability for a student (i)
		$\delta_{1\,i}$	Difference from average student improvement for a student (i)
ϵ	Random Error	ϵ_{ti}	Error associated with student (i) at time (t)

Table 3 Progression of HLMs used to model student performance.

Index	Equation	Parameters
m1	$Y_{ti} = \beta_0 + \delta_{0\,i} + \epsilon_{ti}$	Random Student Intercept
m2	$Y_{ti} = \beta_0 + \beta_1 + \theta_{0\,j(i)} + \theta_{1\,j(i)} + \delta_{0\,i} + \epsilon_{ti}$	Random Student Intercept Nested by Treatment
m3	$Y_{ti} = \beta_0 + \beta_1 + \theta_{0\,j(i)} + \theta_{1\,j(i)} + \delta_{0\,i} + \delta_{1\,i} + \epsilon_{ti}$	Random Student Intercept and Slope Nested by Treatment
m4	$Y_{ti} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \theta_{0\,j(i)} + \theta_{1\,j(i)} + \delta_{0\,i} + \delta_{1\,i} + \epsilon_{ti}$	Random Student and Intercept and Slope Nested by Treatment with Additional Main Effects

The simplest model (m1) attempts to predict student scores based solely on the student's initial performance. This model therefore assumes that no improvement was seen in student scores from week 1 to week 2. While this is clearly not likely, this model was only used as a baseline against which to test the next model.

The second model (m2) builds upon m1 by adding a retest effect and a random effect for the treatment. Within this model, each treatment was allowed to vary in both the intercept ($\theta_{0\,j(i)}$) and slope ($\theta_{1\,j(i)}$). The treatment intercept would help account for any differences in student's initial ability levels between the treatments. This treatment intercept may not be necessary for a study which applies all treatments to a homogeneous sample but because this data collection was conducted at multiple institutions over several years this intercept will help to control for any initial ability level differences that may exist. The treatment slope was allowed to vary because the treatments were expected to cause varying levels of student improvement (this expectation is confirmed in the results section).

The third model (m3) only differs from m2 in that the individual student-level growth is also accounted for in the model. While it is intuitive to recognize that different students will benefit differently from the same treatment, up until this point this variable was not included in the model because it was expected that individual student-level effects would be miniscule compared to the effect caused by the treatment as a whole. Comparing m3 to m2 tests the validity of that assumption. The final model (m4) tests to see if it is beneficial to account for student-level initial ability by using other main effects such as the semester they took the exam and the sex of the students.

The comparison between these models was conducted by using the likelihood ratio test to compare the goodness of fit for each subsequent model (Table 4). The first comparison (m1 to m2), is shown to be significant ($p < 0.001$) which indicates that grouping students by treatment greatly improves the model. When comparing m2 to m3, the goodness of fit between the models is not significant ($p = 0.114$). This confirms the expectation that the individual student-level growth is miniscule compared to the effect caused by the treatment as a whole. The final comparison (m3 to m4) is not significant at the 0.01 level ($p = 0.017$). This comparison is significant at the 0.05 level, but this final model was not used because despite being significant (under this looser requirement), when dealing with a larger sample such as this, even negligible differences can be found to be significantly different. With this in mind, the benefit added to the model by including additional main effects is negligible compared to the inclusion of treatment-effects. Based on these comparisons between the models, all future treatment analysis was conducted using m2 since it was shown that m3 and m4 are not significantly better than m2 and also incurred far greater computational demands. Importantly, the dichotomous coefficients used in the primary manuscript vary slightly from the pilot coefficients. This is because the pilot models were all constructed with the same dataset and therefore had the constraint of needing sex data. Therefore, the final model used in the manuscript had a slightly larger sample size and small changes in the coefficients (n = 1,902).

Table 4 Likelihood ratio test results to compare the goodness of fit between each HLM. Likelihood-ratio and significance correspond to the current model and the previous model (m1 to m2, m2 to m3, and m3 to m4).

Model	Log-Likelihood	Likelihood-Ratio	Significance
m1	-10036.761		
m2	-9848.130	377.261	<0.001
m3	-9845.960	4.341	0.114
m4	-9841.859	8.200	0.017

4) Exam cloning

4.1) Exam recoding

Table 5 shows how the clones of the responses from Exam A were randomized in the creation of Exam B. For example, with question 1: response A was left as response A, response B was moved to response C, response C became response B, and response D was left as response D. This process was carried out through the use of SPSS's recode syntax (IBM Corp, 2017).

Table 5 Explanation of how the exam responses were shuffled to create Exam B.

Question #	Exam A	Exam B	Question #	Exam A	Exam B
Q1	A	A	Q11	A	D
	B	C		B	B
	C	B		C	C
	D	D		D	A
Q2	A	D	Q12	A	A
	B	C		B	C
	C	A		C	B
	D	B		D	D
Q3	A	D	Q13	A	A
	B	C		B	C
	C	A		C	B
	D	B		D	D
Q4	A	C	Q14	A	C
	B	A		B	B
	C	B		C	A
	D	D		D	D
Q5	A	D	Q15	A	A
	B	C		B	B
	C	A		C	D
	D	B		D	C
Q6	A	A	Q16	A	C
	B	C		B	B
	C	D		C	A
	D	B		D	D
Q7	A	B	Q17	A	D
	B	D		B	C
	C	A		C	B
	D	C		D	A
Q8	A	D	Q18	A	B
	B	A		B	C
	C	B		C	D
	D	C		D	A
Q9	A	A	Q19	A	C
	B	B		B	B
	C	D		C	D
	D	C		D	A
Q10	A	A	Q20	A	B
	B	D		B	C
	C	C		C	D
	D	B		D	A

4.2) Exam cloning equivalence

The average week 1 exam performance under each grading method is shown in Table 6. To examine the equivalence between exam clones, exam performance was compared between students (n=2025) who took exam A during week 1 and students (n=219) who took exam B during week 1. By only comparing week 1 performance, there was not yet any feedback intervention. Independent samples t-tests show no significant differences between the exam performances. Similarly, the Cronbach's alphas and the average inter-item correlations are similar as shown in Table 7. Figure 1 shows the performance distributions for each of these grading methods is comparable and that the exam scores are normally distributed. It is also important to note that Exam B had a much smaller sample size than Exam A in week 1, which explains the increase in noise.

These comparisons only showed the exams are comparable in aggregate. To investigate more in-depth, the discrimination and difficulty of each individual question was calculated and compared. Figure 2 shows these values plotted for both exams and shows a spread of difficulty while many items still falling into the range of between 0.3 and 0.8 with discriminations above 0.25 (where harder and easier questions have lower discrimination values). Exact discrimination and difficulty values can be found in Table 8. Item plots to compare test items are shown in Figure 3 and show that the questions are similar for every range of student ability level.

To investigate even deeper than comparing exam items, item answer selections were also compared and are shown in Table 9 and Table 10. These tables show each answer selections, percent selection and attraction indices. Attraction indices were calculated using the top and bottom 25% of students. Green boxes indicate the correct answer for that question and since response options were randomized between Exam A and Exam B, values should not be directly compared between the tables without realignment.

Table 6 Comparison of week 1 exam performances showing no significant differences.

	Dichotomous		Open		Hierarchy	
	Exam A	Exam B	Exam A	Exam B	Exam A	Exam B
Mean	12.56	12.34	13.75	13.62	14.55	14.35
Std Dev	4.12	4.28	3.57	3.70	3.16	3.33
n	2025	219	2025	219	2025	219
t(p)	0.740 (0.460)		0.463 (0.644)		0.869 (0.385)	
Cohen's d	0.052		0.036		0.062	

Table 7 Comparison of exam Cronbach's alphas and average inter-item correlations.

	Dichotomous		Hierarchy		Open	
	Exam A	Exam B	Exam A	Exam B	Exam A	Exam B
Cronbach's Alpha	0.787	0.807	0.785	0.805	0.789	0.809
Average Inter-Item Correlation	0.154	0.171	0.153	0.170	0.156	0.173

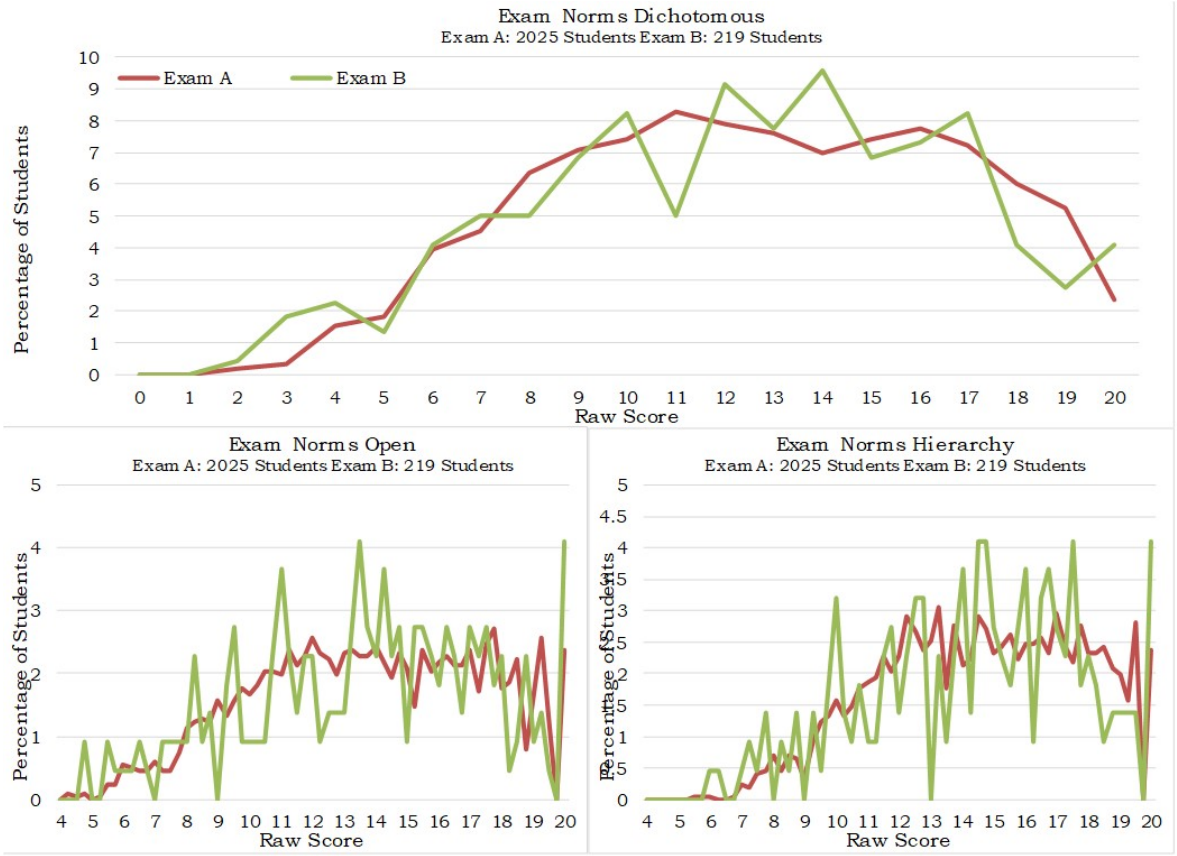


Fig. 1 Percent of each raw score obtained on Exam A and Exam B under each of the grading methods.

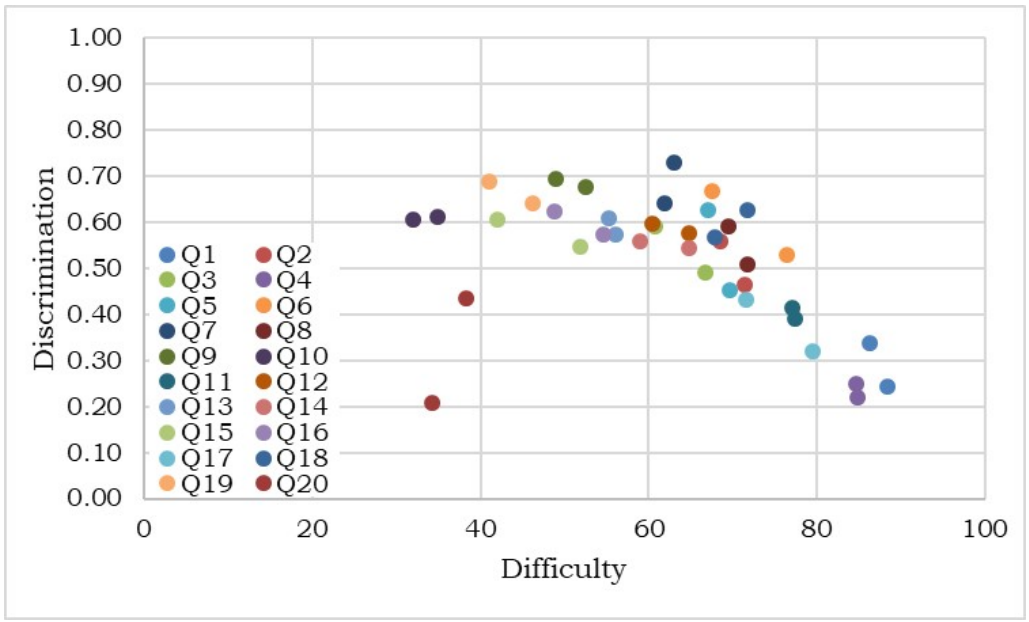
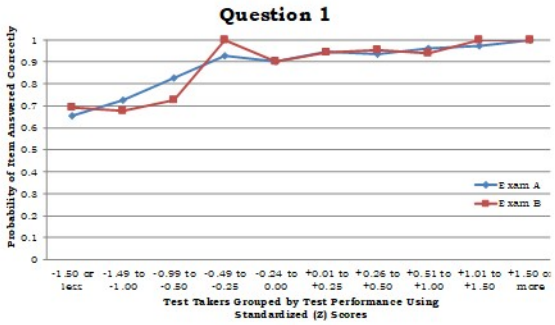
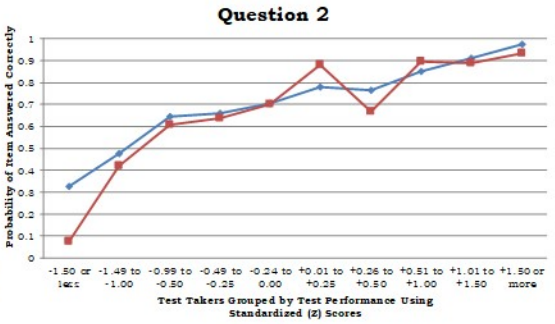


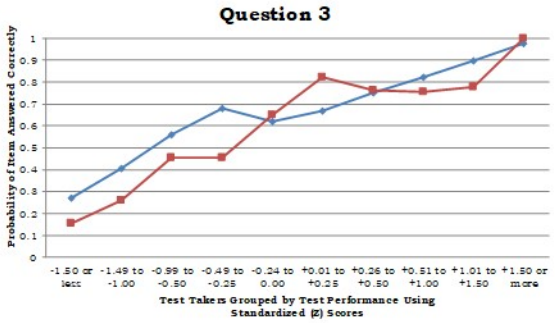
Fig. 2 Plot of item difficulty versus discrimination. Each question has two points, one from exam A and one from exam B.

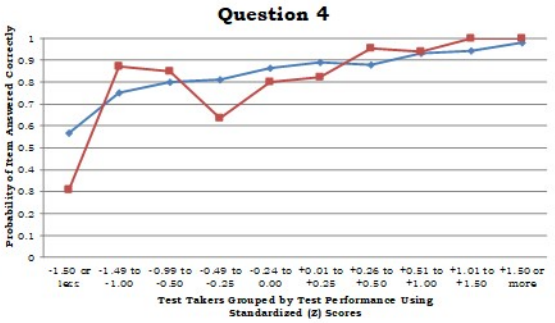
Table 8 Item difficulty and discrimination for each exam.

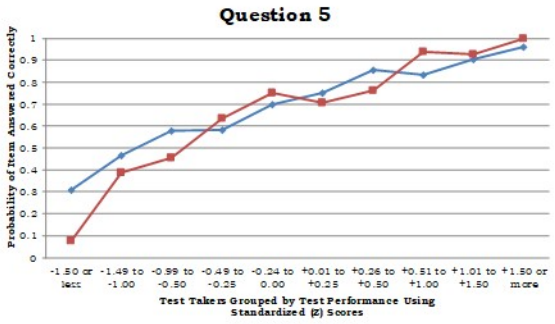
Exam A		Exam B	
Difficulty	**Discrimination**	**Difficulty**	**Discrimination**
88.44	0.24	86.30	0.34
71.41	0.46	68.49	0.56
66.81	0.49	60.73	0.59
84.69	0.25	84.93	0.22
69.68	0.45	67.12	0.63
76.40	0.53	67.58	0.67
61.83	0.64	63.01	0.73
69.58	0.59	71.69	0.51
48.94	0.69	52.51	0.68
34.91	0.61	31.96	0.61
77.48	0.39	77.17	0.41
60.40	0.60	64.84	0.58
56.15	0.57	55.25	0.61
59.01	0.54	64.84	0.56
51.85	0.55	42.01	0.61
54.57	0.57	48.86	0.62
71.65	0.43	79.45	0.32
67.95	0.63	71.69	0.57
46.17	0.64	41.10	0.69
38.32	0.43	34.25	0.21

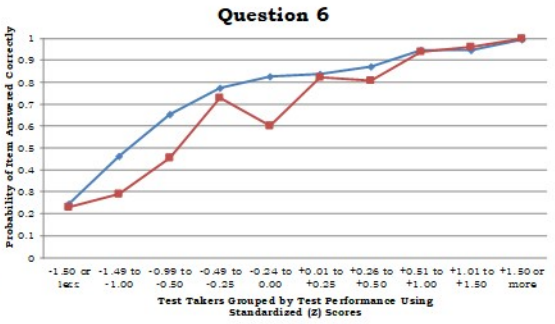

Question 1
Probability of Item Answered Correctly
Exam A
Exam B
-1.50 or less
-1.49 to -1.00
-0.99 to -0.50
-0.49 to -0.25
-0.24 to 0.00
+0.01 to +0.25
+0.26 to +0.50
+0.51 to +1.00
+1.01 to +1.50
+1.50 or more
Test Takers Grouped by Test Performance Using Standardized (Z) Scores

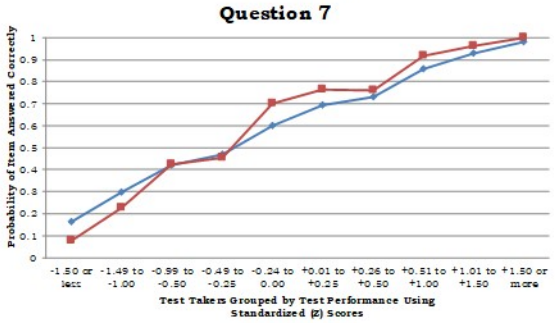

Question 2
Probability of Item Answered Correctly
Test Takers Grouped by Test Performance Using Standardized (Z) Scores

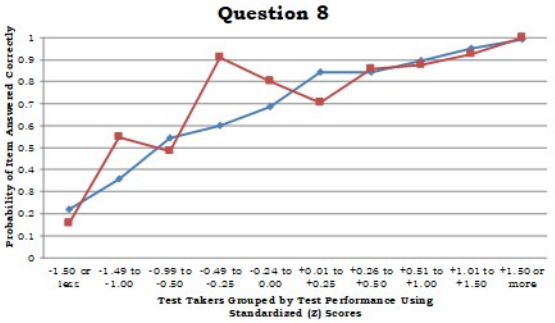

Question 3
Probability of Item Answered Correctly
Test Takers Grouped by Test Performance Using Standardized (Z) Scores

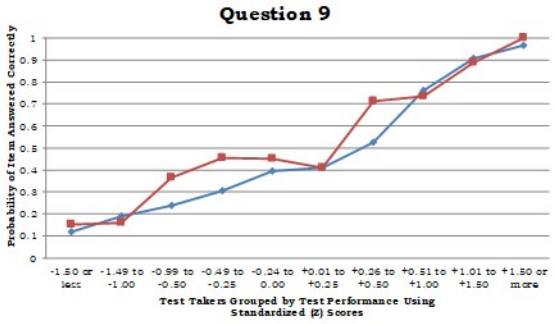

Question 4
Probability of Item Answered Correctly
Test Takers Grouped by Test Performance Using Standardized (Z) Scores

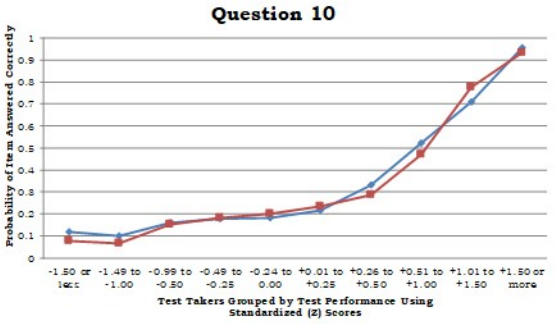

Question 5
Probability of Item Answered Correctly
Test Takers Grouped by Test Performance Using Standardized (Z) Scores


Question 6
Probability of Item Answered Correctly
Test Takers Grouped by Test Performance Using Standardized (Z) Scores


Question 7
Probability of Item Answered Correctly
Test Takers Grouped by Test Performance Using Standardized (Z) Scores


Question 8
Probability of Item Answered Correctly
Test Takers Grouped by Test Performance Using Standardized (Z) Scores


Question 9
Probability of Item Answered Correctly
Test Takers Grouped by Test Performance Using Standardized (Z) Scores


Question 10
Probability of Item Answered Correctly
Test Takers Grouped by Test Performance Using Standardized (Z) Scores

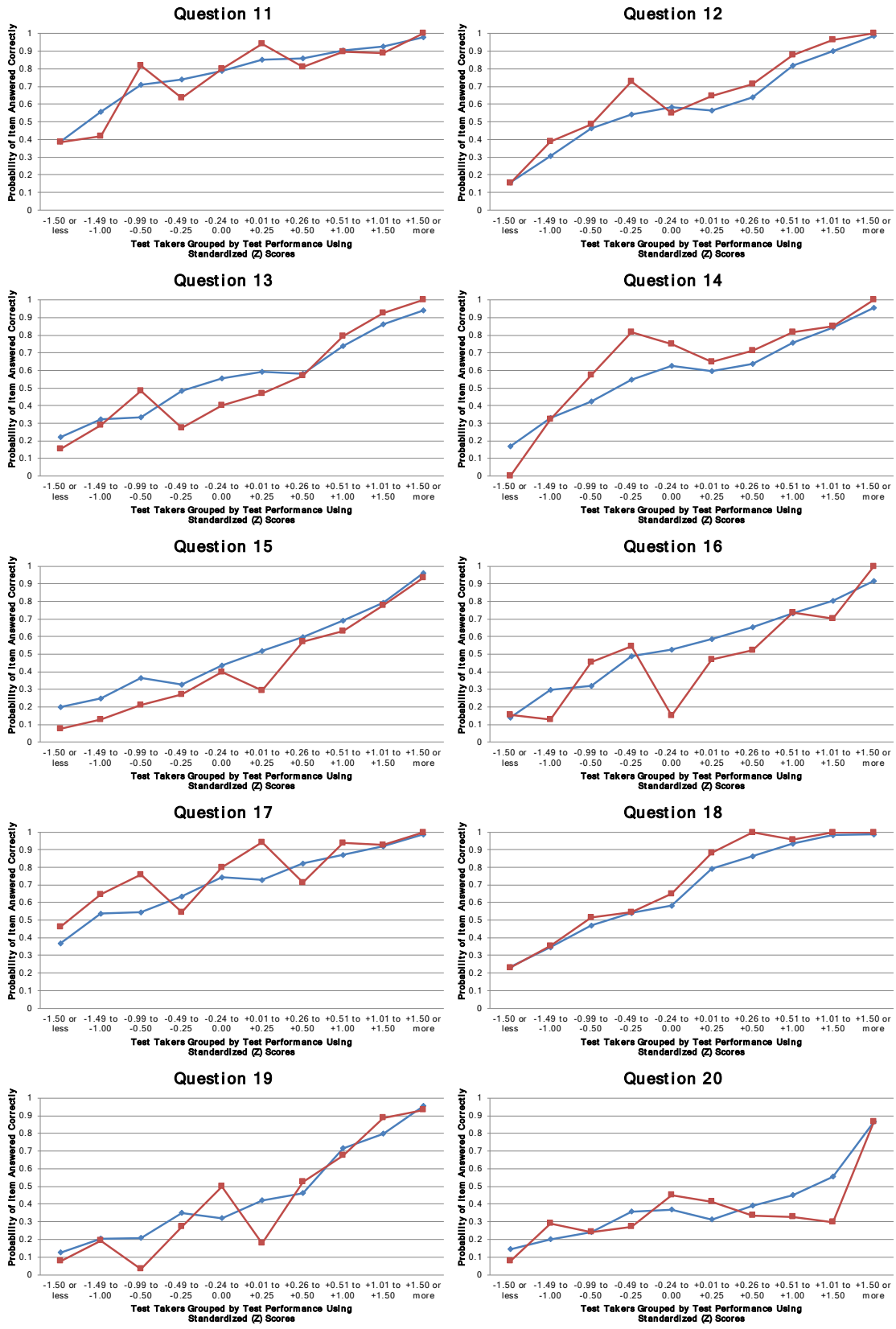


Fig. 3 Item plots of each question for Exam A (blue) and Exam B (red) showing similar performance. Item plots were constructed in the same manner as discussed by Holme (Holme and Murphy, 2011).

Table 9 The percent of students who chose each response option (e.g., for Q1 0.79% of students selected response A), and the attraction indices for each response within Exam A.

Exam A	%A	%B	%C	%D	Attraction A	Attraction B	Attraction C	Attraction D
Q1	0.79%	2.02%	88.44%	8.74%	-0.02	-0.03	0.24	-0.19
Q2	10.62%	8.35%	9.58%	71.41%	-0.09	-0.20	-0.17	0.46
Q3	66.81%	11.36%	7.21%	14.57%	0.49	-0.20	-0.12	-0.17
Q4	84.69%	5.09%	2.37%	7.80%	0.25	-0.05	-0.03	-0.17
Q5	3.90%	20.79%	69.68%	5.63%	-0.08	-0.28	0.45	-0.09
Q6	11.21%	4.44%	7.95%	76.40%	-0.28	-0.12	-0.13	0.53
Q7	14.47%	17.38%	6.27%	61.83%	-0.25	-0.26	-0.13	0.64
Q8	8.49%	69.58%	16.44%	5.38%	-0.16	0.59	-0.33	-0.10
Q9	48.94%	30.86%	15.11%	5.04%	0.69	-0.33	-0.27	-0.10
Q10	15.60%	29.19%	34.91%	20.15%	-0.09	-0.27	0.61	-0.25
Q11	77.48%	4.35%	7.95%	10.22%	0.39	-0.06	-0.12	-0.20
Q12	2.32%	4.49%	32.74%	60.40%	-0.04	-0.07	-0.49	0.60
Q13	19.16%	56.15%	16.05%	8.54%	-0.23	0.57	-0.21	-0.13
Q14	6.52%	59.01%	24.89%	9.48%	-0.10	0.54	-0.22	-0.22
Q15	51.85%	15.16%	12.00%	20.89%	0.55	-0.20	-0.11	-0.24
Q16	30.12%	7.65%	7.51%	54.57%	-0.28	-0.17	-0.12	0.57
Q17	6.72%	71.65%	4.10%	17.38%	-0.11	0.43	-0.11	-0.21
Q18	8.15%	67.95%	15.41%	8.44%	-0.16	0.63	-0.28	-0.18
Q19	31.95%	9.48%	46.17%	12.20%	-0.43	-0.16	0.64	-0.05
Q20	38.32%	13.04%	19.90%	28.49%	0.43	-0.07	-0.20	-0.17

Table 10 The percent of students who chose each response option (e.g., for Q1 0.91% of students selected response A), and the attraction indices for each response within Exam B.

Exam B	%A	%B	%C	%D	Attraction A	Attraction B	Attraction C	Attraction D
Q1	0.91%	86.30%	1.83%	10.96%	-0.02	0.34	-0.03	-0.25
Q2	8.68%	68.49%	9.13%	13.70%	-0.22	0.56	-0.24	-0.09
Q3	6.85%	18.26%	13.70%	60.73%	-0.10	-0.27	-0.20	0.53
Q4	2.74%	2.74%	84.93%	9.59%	-0.03	0.00	0.20	-0.17
Q5	67.12%	10.05%	18.26%	4.57%	0.63	-0.27	-0.25	-0.08
Q6	21.00%	67.58%	7.76%	2.74%	-0.49	0.65	-0.18	-0.02
Q7	8.68%	12.79%	63.01%	15.53%	-0.22	-0.22	0.73	-0.31
Q8	71.69%	7.31%	14.16%	6.85%	0.51	-0.08	-0.27	-0.14
Q9	52.51%	23.74%	17.81%	5.94%	0.68	-0.27	-0.39	-0.05
Q10	19.18%	23.29%	31.96%	25.57%	-0.17	-0.32	0.63	-0.13
Q11	15.53%	2.74%	4.11%	77.17%	-0.28	-0.05	-0.09	0.38
Q12	1.37%	29.22%	4.57%	64.84%	-0.03	-0.49	-0.07	0.59
Q13	19.18%	12.79%	55.25%	12.79%	-0.17	-0.15	0.61	-0.32
Q14	19.18%	64.84%	5.94%	10.05%	-0.29	0.54	-0.10	-0.17
Q15	42.01%	15.07%	30.14%	12.33%	0.61	-0.15	-0.26	-0.15
Q16	3.65%	6.85%	40.18%	48.86%	-0.05	-0.17	-0.36	0.61
Q17	3.20%	4.57%	79.45%	12.33%	-0.02	-0.10	0.32	-0.18
Q18	12.33%	6.85%	71.69%	8.22%	-0.24	-0.12	0.55	-0.19
Q19	13.70%	7.76%	36.53%	41.10%	-0.09	-0.05	-0.53	0.68
Q20	35.62%	34.25%	12.33%	16.89%	0.05	0.21	-0.06	-0.15

Besides the quantitative comparisons shown above, the exams were also compared using multimode scoring. A brief description of multimode is included in the introduction but precise details about how multimode scores were calculated can be found in the previous work that has been done on these exams (Murphy et al.). These estimates were conducted based on raters' expectations of response patterns for each ability so students who had response patterns that were not predicted were placed into an "other" category. Sankey diagrams for student categorization and movement between the content areas are shown in Figure 4 and Figure 5. The populations of each categorization and movements between them are similar for each content area. Based on how often a student was categorized into each ability level, and what ability levels they fell into, the student's overall ability was estimated. Again, the specifics of the methods followed to achieve this overall ability estimate can be found in previous work (Murphy et al.). The overall ability level distributions were shown to be comparable and are visualized in Figure 6.

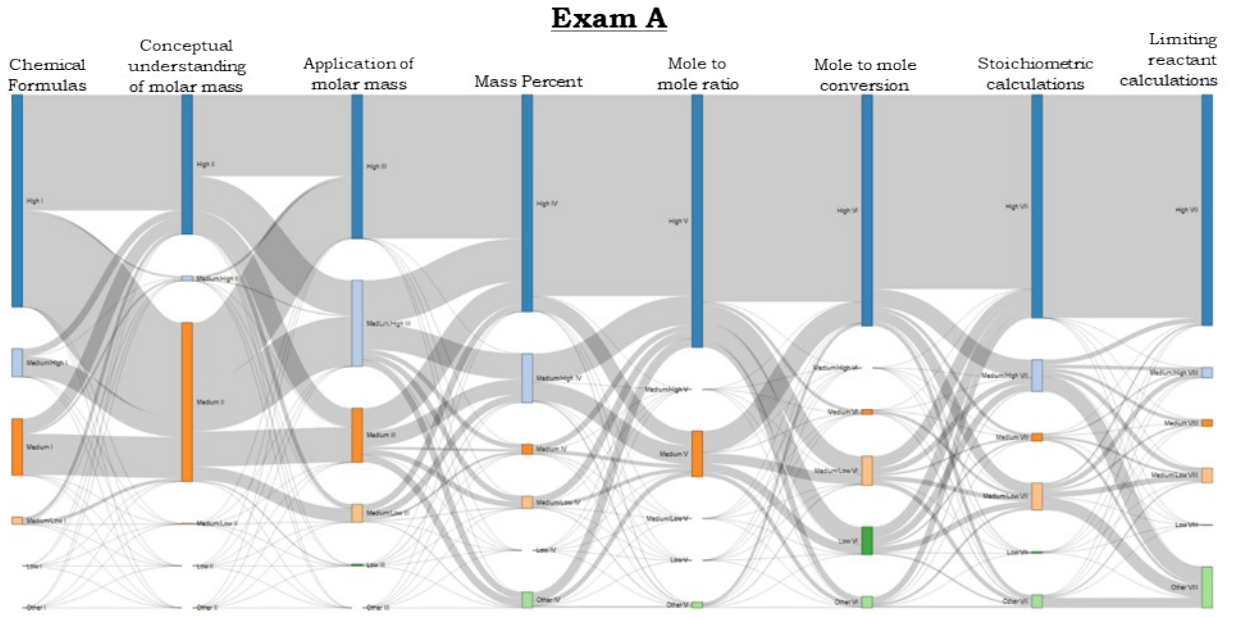


Fig. 4 Sankey diagram for exam A showing student categorization (high, medium/high, medium, medium/low, low, or other) and movement between predicted ability levels within content areas.

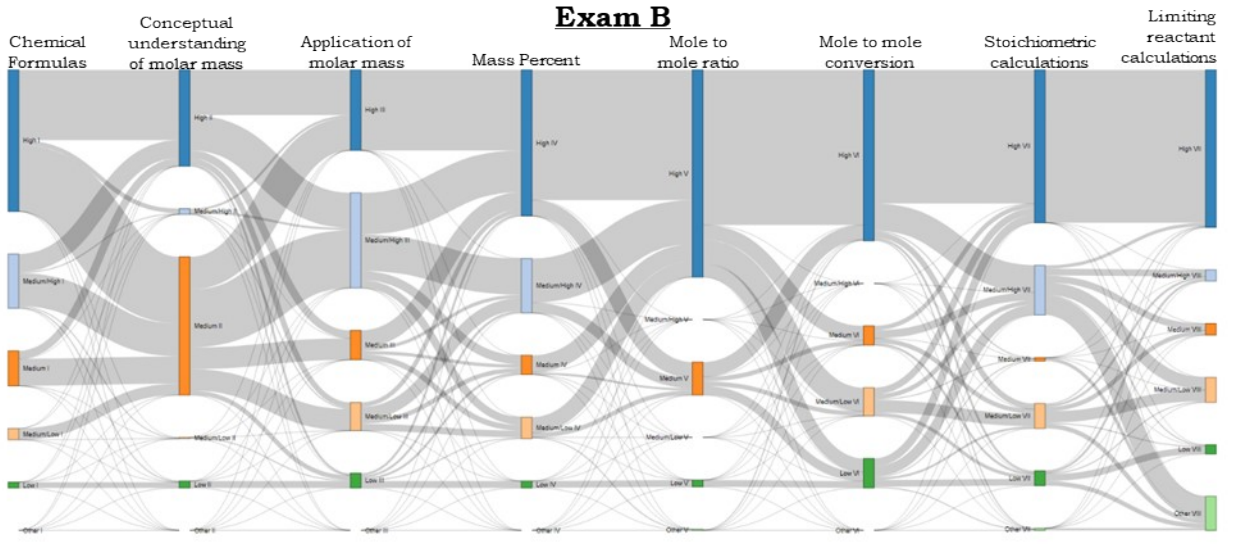


Fig. 5 Sankey diagram for exam B showing student categorization (high, medium/high, medium, medium/low, low, or other) and movement between predicted ability levels within content areas.

Fig. 6 Predicted overall ability of students based on multimode results.

5) Pilot model coefficients

Table 11 The equations and coefficients of the pilot models that were used to determine which model would be most appropriate. These models were all built with 1,898 students because 4 students had to be removed because of missing sex data. Student-level intercepts ($\delta_{0\,i}$) and slopes ($\delta_{1\,i}$) also generated but are not included for brevity and irrelevance to the research question. The dummy coding for m4 is as follows: Sex: female = 0 and male=1, Semester: Fall = 0 and Spring = 1.

m1	$Y_{ti} = \beta_0 + \epsilon_{ti}$											
	β_0	13.722										
m2	$Y_{ti} = \beta_0 + \beta_1 + \theta_{0\,j(i)} + \theta_{1\,j(i)} + \delta_{0\,i} + \epsilon_{ti}$											
	β_0	13.433										
	β_1	1.260										
	Treatment	1	2	3	4	5	6	7	8	9	10	11
	$\theta_{0\,j(i)}$	-0.716	0.315	-0.753	-0.529	-0.343	1.426	0.431	1.085	-0.018	-0.224	-0.675
	$\theta_{1\,j(i)}$	-0.919	-0.227	-0.491	0.274	0.326	0.074	0.014	0.574	-0.387	0.840	-0.077
m3	$Y_{ti} = \beta_0 + \beta_1 + \theta_{0\,j(i)} + \theta_{1\,j(i)} + \delta_{0\,i} + \delta_{1\,i} + \epsilon_{ti}$											
	β_0	13.429										
	β_1	1.260										
	Treatment	1	2	3	4	5	6	7	8	9	10	11
	$\theta_{0\,j(i)}$	-0.710	0.315	-0.736	-0.516	-0.328	1.378	0.424	1.058	-0.017	-0.213	-0.655
	$\theta_{1\,j(i)}$	-0.915	-0.223	-0.484	0.269	0.315	0.081	0.017	0.564	-0.374	0.829	-0.078
m4	$Y_{ti} = \beta_0 + \beta_1 + \beta_2 + \beta_3 + \theta_{0\,j(i)} + \theta_{1\,j(i)} + \delta_{0\,i} + \delta_{1\,i} + \epsilon_{ti}$											
	β_0	13.312										
	β_1	1.259										
	β_2	-0.107										
	β_3	0.538										
	Treatment	1	2	3	4	5	6	7	8	9	10	11
	$\theta_{0\,j(i)}$	-0.697	0.304	-0.754	-0.485	-0.372	1.422	0.397	1.072	-0.072	-0.171	-0.642
	$\theta_{1\,j(i)}$	-0.914	-0.223	-0.485	0.269	0.309	0.090	0.015	0.566	-0.379	0.830	-0.078

5.1) Investigation into Q20 removal

Interviews with students revealed that question 20 may have been misinterpreted by some students. This misunderstanding may be the root cause for why the questions' discrimination was not consistent between the exams. Later IRT analysis also confirmed inconsistent and poor discrimination of this question. Because of the weakness of this question, an investigation was conducted to determine if removal of this question from analysis would be appropriate. To test this, m2 was constructed for each grading scheme both with and without Q20 and the models were compared. All of these models were built with the full sample of 1,902 students for which week 1 and week 2 data was available. While a direct comparison between coefficients can be conducted (Table 12 compared to Table 13), it is of limited value. The reason for this limitation can be seen for example when comparing the dichotomous models. The mean slope (β_1) for the model including Q20 is larger than the model without Q20. However, the model with Q20 accounts for some this difference by having a more negative treatment-level slope ($\theta_{1\,j(i)}$). In other words, often when the mean was larger the amount to subtract from that mean was also greater so comparing just raw coefficients leads to differences being maximized between the models.

This issue can be circumvented by comparing the direct amount each treatment benefited ($\beta_1 + \theta_{1\,j(i)}$) as opposed to the amount away ($\theta_{1\,j(i)}$) from an estimated average (β_1). These values are shown in Table 14 and Table 15 and show similar results. The growths along with the standard error are plotted in Figure 7 through Figure 9 and show overlap of every treatment under every grading scheme. Seeing no significant difference between the coefficients with and without the Q20 the question was not removed. This decision was further confirmed when analyzing the model fits and seeing relatively minor differences (Table 15).

Table 12 Model coefficients when including Q20.

$Y_{ti} = \beta_0 + \beta_1 + \theta_{0\,j(i)} + \theta_{1\,j(i)} + \delta_{0\,i} + \epsilon_{ti}$												
Dichotomous	β_0	13.432										
	β_1	1.259										
	Treatment	1	2	3	4	5	6	7	8	9	10	11
	$\theta_{0\,j(i)}$	-0.718	0.308	-0.751	-0.528	-0.343	1.427	0.432	1.086	-0.016	-0.223	-0.673
	$\theta_{1\,j(i)}$	-0.921	-0.237	-0.491	0.275	0.327	0.075	0.015	0.576	-0.386	0.842	-0.076
Open	β_0	14.494										
	β_1	1.103										
	Treatment	1	2	3	4	5	6	7	8	9	10	11
	$\theta_{0\,j(i)}$	-0.561	0.287	-0.661	-0.491	-0.329	1.230	0.378	0.967	-0.053	-0.200	-0.567
	$\theta_{1\,j(i)}$	-0.786	-0.222	-0.343	0.305	0.312	0.032	-0.034	0.446	-0.336	0.707	-0.081
Hierarchy	β_0	15.203										
	β_1	0.975										
	Treatment	1	2	3	4	5	6	7	8	9	10	11
	$\theta_{0\,j(i)}$	-0.504	0.216	-0.625	-0.426	-0.188	1.027	0.307	0.801	0.028	-0.122	-0.514
	$\theta_{1\,j(i)}$	-0.721	-0.194	-0.281	0.283	0.239	0.089	0.021	0.390	-0.325	0.535	-0.035

Table 13 Model coefficients when removing Q20.

$Y_{ti} = \beta_0 + \beta_1 + \theta_{0\,j(i)} + \theta_{1\,j(i)} + \delta_{0\,i} + \epsilon_{ti}$												
Dichotomous	β_0	13.017										
	β_1	1.133										
	Treatment	1	2	3	4	5	6	7	8	9	10	11
	$\theta_{0\,j(i)}$	-0.671	0.319	-0.664	-0.523	-0.315	1.357	0.393	0.974	0.004	-0.271	-0.603
	$\theta_{1\,j(i)}$	-0.852	-0.250	-0.438	0.235	0.275	0.104	-0.010	0.529	-0.325	0.830	-0.097
Open	β_0	13.977										
	β_1	0.988										
	Treatment	1	2	3	4	5	6	7	8	9	10	11
	$\theta_{0\,j(i)}$	-0.521	0.290	-0.590	-0.487	-0.310	1.178	0.346	0.876	-0.042	-0.231	-0.509
	$\theta_{1\,j(i)}$	-0.719	-0.218	-0.290	0.267	0.272	0.041	-0.060	0.408	-0.290	0.685	-0.097
Hierarchy	β_0	14.657										
	β_1	0.844										
	Treatment	1	2	3	4	5	6	7	8	9	10	11
	$\theta_{0\,j(i)}$	-0.472	0.223	-0.559	-0.425	-0.174	0.990	0.277	0.726	0.025	-0.150	-0.461
	$\theta_{1\,j(i)}$	-0.633	-0.184	-0.237	0.241	0.199	0.075	-0.003	0.357	-0.282	0.509	-0.042

Table 14 Treatment initial ability and growth when including Q20.

	Treatment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
Dichotomous	$\beta_0 + \theta_{0\,j(i)}$	12.713	13.739	12.680	12.904	13.089	14.858	13.864	14.518	13.415	13.208	12.758
	$\beta_1 + \theta_{1\,j(i)}$	0.337	1.021	0.768	1.534	1.586	1.333	1.274	1.835	0.873	2.101	1.183
Open	$\beta_0 + \theta_{0\,j(i)}$	13.934	14.781	13.833	14.003	14.165	15.725	14.872	15.461	14.441	14.294	13.928
	$\beta_1 + \theta_{1\,j(i)}$	0.317	0.881	0.760	1.407	1.415	1.135	1.068	1.549	0.766	1.810	1.021
Hierarchy	$\beta_0 + \theta_{0\,j(i)}$	14.699	15.418	14.577	14.777	15.015	16.230	15.510	16.003	15.230	15.080	14.689
	$\beta_1 + \theta_{1\,j(i)}$	0.254	0.781	0.694	1.258	1.214	1.064	0.996	1.365	0.650	1.510	0.940

Table 15 Treatment initial ability and growth when removing Q20.

	Treatment	1	2	3	4	5	6	7	8	9	10	11
Dichotomous	$\beta_0 + \theta_{0\,j(i)}$	12.347	13.336	12.353	12.494	12.703	14.374	13.411	13.991	13.022	12.747	12.415
	$\beta_1 + \theta_{1\,j(i)}$	0.281	0.883	0.696	1.368	1.408	1.237	1.123	1.662	0.808	1.963	1.037
Open	$\beta_0 + \theta_{0\,j(i)}$	13.456	14.267	13.387	13.490	13.667	15.156	14.323	14.853	13.935	13.746	13.468
	$\beta_1 + \theta_{1\,j(i)}$	0.269	0.770	0.698	1.254	1.259	1.029	0.928	1.396	0.698	1.673	0.891
Hierarchy	$\beta_0 + \theta_{0\,j(i)}$	14.185	14.879	14.098	14.232	14.482	15.646	14.934	15.382	14.682	14.507	14.196
	$\beta_1 + \theta_{1\,j(i)}$	0.211	0.660	0.607	1.085	1.043	0.919	0.840	1.201	0.562	1.353	0.802

Table 16 Model fits with and without Q20 for each of the grading schemes.

	Log Likelihood with Q20	Log Likelihood Without Q20
Dichotomous	-9868.134	-9729.839
Open	-9275.585	-9143.314
Hierarchy	-8835.852	-8688.663

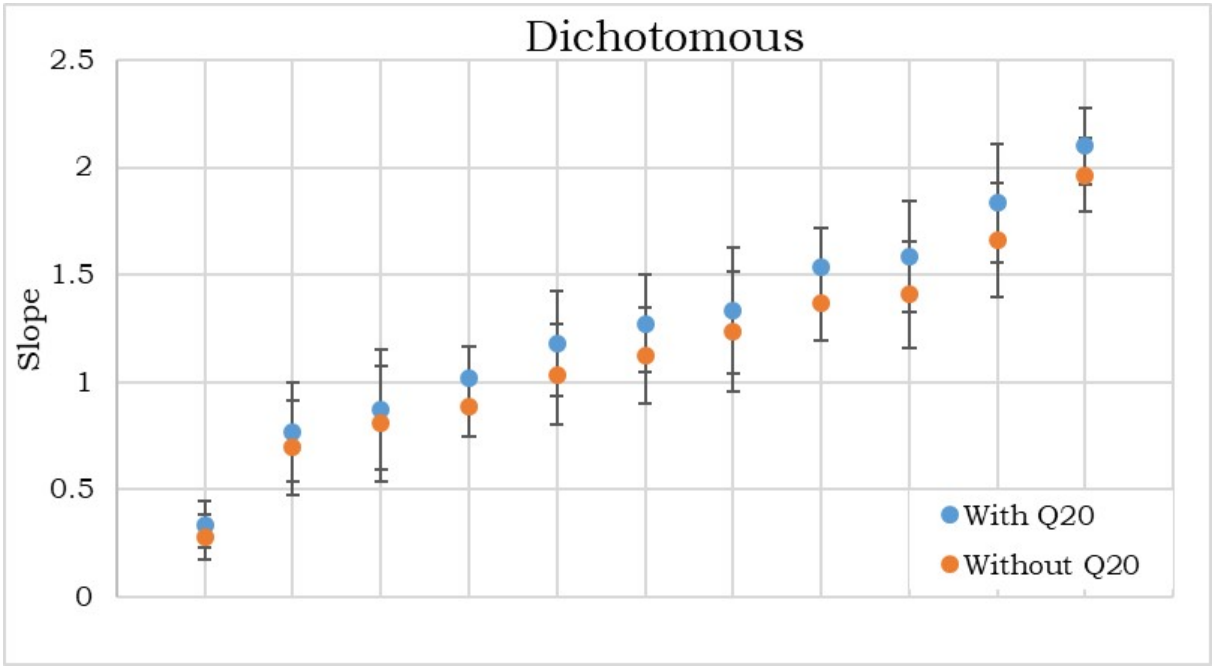


Fig. 7 Dichotomous student growth, along with the standard error, caused by each treatment both with and without Q20.

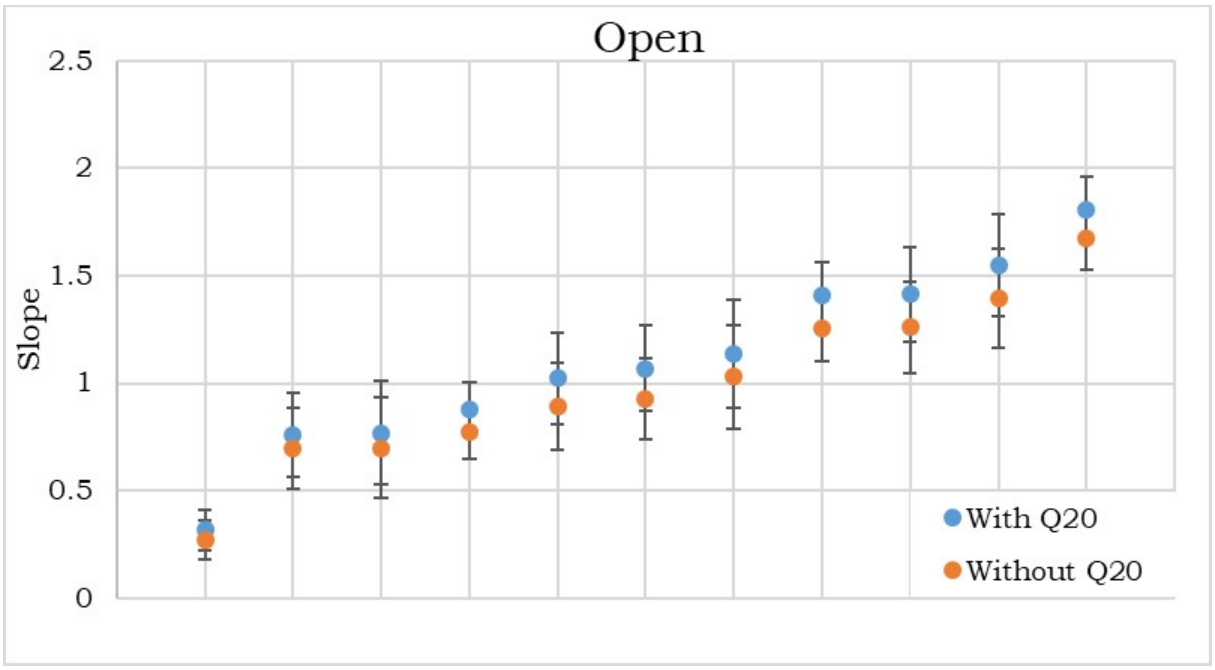


Fig 8 Open student growth, along with the standard error, caused by each treatment both with and without Q20.

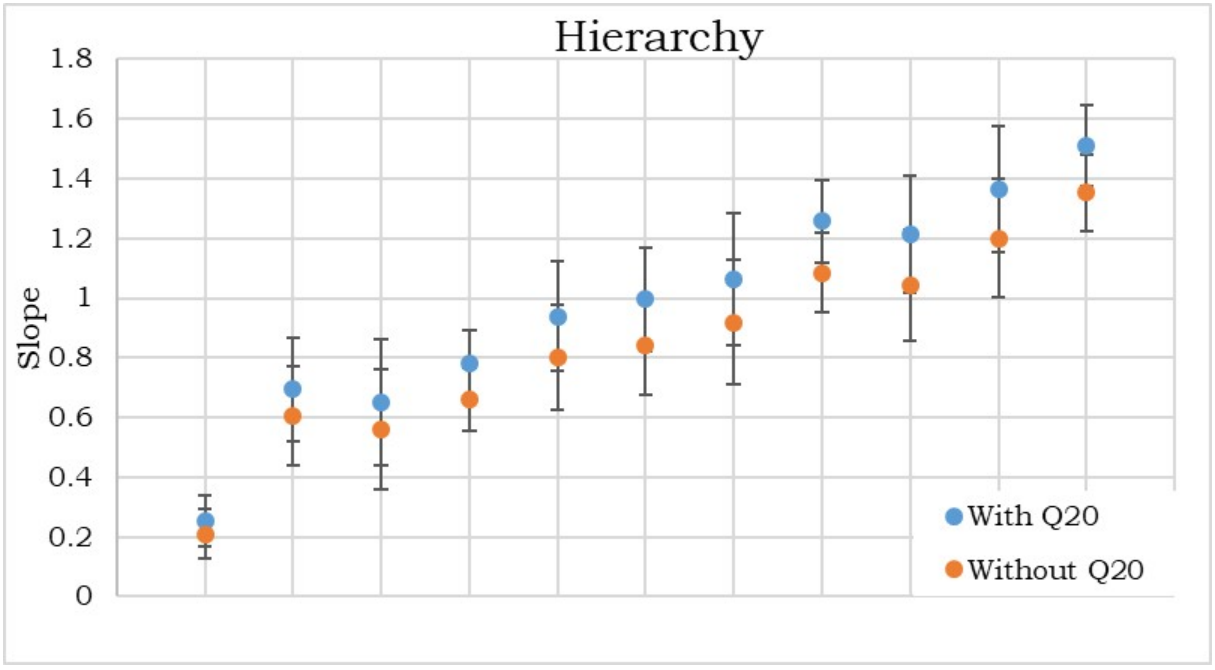


Fig. 8 Hierarchy student growth, along with the standard error, caused by each treatment both with and without Q20.

6) Sample conceptual item feedback

4) Which sample contains the LARGEST mass?
   a. 1.0 mol of NO_2
   b. 1.0 mol N_2O
   c. 1.0 mol NO
   d. All would have the same mass because they all contain the same moles

1.0 mol NO is incorrect. To arrive at this answer it is likely that you chose the sample with the smallest molar mass.

The correct answer is 1.0 mol NO2 will have the largest mass.

Using the generic formula:

Mass of sample = Moles of sample x Molar mass of sample

As all 3 samples contain the same number of moles, the sample with the largest molar mass will also have the largest mass.

Fig. 9 Example of feedback given to a student who incorrectly selected response "c."

7) Treatment slopes ordered by quantity of feedback

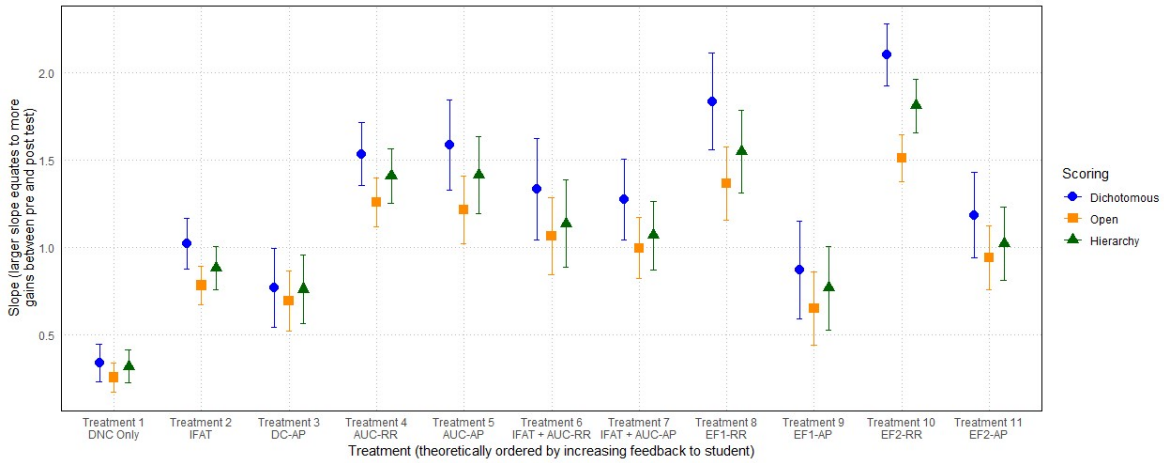


Fig. 10 Estimates for the slope ($\beta_1 + \theta_{1\,j(i)}$) of each treatment under the m2 model. The slope of each treatment is interpreted as how many points of improvement were caused by that treatment. Error bars correspond to the standard error of the treatment slope. Plot is functionally identical to Fig. 1 in the main text though is now ordered by the quantity of feedback provided to students.

8) Treatment grouping coefficients

Table 17 Model coefficients for the modified m2 model after samples which received similar treatments were collapsed into their 4 groupings.

$Y_{ti} = \beta_0 + \beta_1 + \theta_{0\,k(i)} + \theta_{1\,k(i)} + \delta_{0\,i} + \epsilon_{ti}$					
Dichotomous	β_0	13.275			
	β_1	1.206			
	Treatment Group	1	2	3	4
	$\theta_{0\,k(i)}$	0.406	0.135	-0.092	-0.449
	$\theta_{1\,k(i)}$	0.840	0.280	-0.191	-0.928
Open	β_0	14.369			
	β_1	1.059			
	Treatment Group	1	2	3	4
	$\theta_{0\,k(i)}$	0.314	0.112	-0.075	-0.351
	$\theta_{1\,k(i)}$	0.707	0.251	-0.169	-0.789
Hierarchy	β_0	15.089			
	β_1	0.921			
	Treatment Group	1	2	3	4
	$\theta_{0\,k(i)}$	0.275	0.121	-0.063	-0.333
	$\theta_{1\,k(i)}$	0.581	0.256	-0.134	-0.703
True Score	β_0	15.021			
	β_1	1.188			
	Treatment Group	1	2	3	4
	$\theta_{0\,k(i)}$	0.539	0.365	0.022	-0.926
	$\theta_{1\,k(i)}$	0.518	0.351	0.021	-0.889

Table 18 Treatment grouping initial ability and growth.

	Treatment Group	1	2	3	4
Dichotomous	$\beta_0 + \theta_{0\,k(i)}$	13.681	13.411	13.183	12.827
	$\beta_1 + \theta_{1\,k(i)}$	2.045	1.486	1.014	0.277
Open	$\beta_0 + \theta_{0\,k(i)}$	14.683	14.481	14.294	14.018
	$\beta_1 + \theta_{1\,k(i)}$	1.766	1.310	0.890	0.270
Hierarchy	$\beta_0 + \theta_{0\,k(i)}$	15.364	15.210	15.026	14.757
	$\beta_1 + \theta_{1\,k(i)}$	1.502	1.177	0.787	0.217
True Score	$\beta_0 + \theta_{0\,k(i)}$	15.560	15.386	15.042	14.095
	$\beta_1 + \theta_{1\,k(i)}$	1.706	1.539	1.209	0.299

Table 19 Log likelihood of treatment grouping under each grading scheme.

	Log Likelihood
Dichotomous	-9872.201
Open	-9279.716
Hierarchy	-8838.57
True Score	-10707.635

9) Item response theory results

Table 20 Difficulty and discrimination as calculated by IRT for each of the treatment groupings.

	Treatment Grouping 1				Treatment Grouping 2			
	Week 1		Week 2		Week 1		Week 2	
Question	Difficulty	Discrimination	Difficulty	Discrimination	Difficulty	Discrimination	Difficulty	Discrimination
Q1	-2.152	1.083	-7.593	0.472	-2.616	1.030	-5.943	0.590
Q2	-2.034	0.651	-2.049	0.997	-2.390	0.512	-2.047	0.872
Q3	-1.145	1.103	-1.391	1.094	-1.233	0.942	-1.992	0.806
Q4	-3.795	0.391	-4.010	1.060	-2.887	0.725	-4.296	0.682
Q5	-1.596	0.861	-1.243	1.894	-1.166	1.071	-1.290	1.565
Q6	-1.956	0.871	-1.613	1.711	-1.531	1.440	-1.624	1.840
Q7	-0.618	1.512	-0.525	1.691	-0.847	1.126	-0.699	1.534
Q8	-1.262	1.716	-1.203	2.552	-1.010	1.626	-0.904	2.692
Q9	-0.147	1.387	-0.636	2.461	-0.285	1.753	-0.360	1.893
Q10	0.438	1.431	-0.690	1.085	0.323	1.419	-0.080	1.690
Q11	-3.453	0.440	-2.264	1.115	-1.920	0.718	-1.903	0.877
Q12	-0.682	1.433	-0.990	2.014	-0.571	1.006	-1.013	1.015
Q13	-0.170	1.048	-1.173	1.286	-0.350	0.993	-1.375	1.044
Q14	-0.428	1.073	-0.951	1.317	-0.808	0.711	-0.915	1.142
Q15	-0.413	0.992	-0.257	1.600	-0.495	0.927	-0.405	1.308
Q16	-0.593	0.919	-1.277	1.476	-0.500	0.881	-1.053	1.591
Q17	-1.744	0.892	-2.057	1.164	-1.425	0.978	-1.885	0.974
Q18	-0.957	1.607	-0.987	2.215	-0.803	1.821	-0.867	2.276
Q19	-0.055	1.849	-0.859	1.592	-0.121	1.146	-0.647	1.351
Q20	0.097	0.561	-0.499	1.401	0.473	0.706	-0.584	0.622

	Treatment Grouping 3				Treatment Grouping 4			
	Week 1		Week 2		Week 1		Week 2	
Question	Difficulty	Discrimination	Difficulty	Discrimination	Difficulty	Discrimination	Difficulty	Discrimination
Q1	-3.553	0.779	-4.224	0.819	-2.331	0.987	-2.708	0.915
Q2	-1.810	0.745	-2.326	0.760	-1.477	0.731	-1.341	0.989
Q3	-1.191	0.677	-1.620	0.747	-0.935	0.964	-1.103	0.823
Q4	-2.120	1.173	-4.724	0.546	-2.329	0.855	-2.411	0.841
Q5	-1.024	1.266	-1.199	1.295	-1.084	0.920	-1.013	1.016
Q6	-1.266	1.676	-1.741	1.438	-1.020	1.623	-1.034	2.165
Q7	-0.543	1.565	-0.628	1.539	-0.536	1.379	-0.470	2.027
Q8	-1.044	1.612	-0.900	1.822	-0.739	1.560	-0.826	1.386
Q9	-0.177	1.329	-0.262	1.743	0.113	1.261	-0.016	1.366
Q10	0.332	1.553	0.134	1.672	0.789	1.292	0.505	1.423
Q11	-2.380	0.688	-1.667	1.042	-1.650	1.079	-1.394	1.210
Q12	-0.606	1.145	-0.897	0.926	-0.652	1.257	-0.546	1.343
Q13	-0.403	0.832	-1.016	1.298	-0.616	1.077	-0.583	1.302
Q14	-0.752	0.858	-1.003	0.952	-0.494	1.139	-0.627	1.143
Q15	-0.319	0.831	-0.193	0.973	0.073	0.930	0.096	1.123
Q16	-0.593	0.952	-0.994	0.926	-0.254	0.936	-0.332	0.973
Q17	-1.402	0.901	-2.083	0.943	-1.335	1.101	-1.605	1.313
Q18	-0.775	2.040	-0.879	2.013	-0.737	1.856	-0.538	2.047
Q19	-0.226	1.218	-0.297	1.505	0.149	1.059	-0.049	1.216
Q20	0.984	0.558	0.023	0.756	1.034	0.556	0.418	0.825

10) Lord results

Table 21 Significance values from Lord's statistic for DIF between week 1 and week 2. Values below 0.001 are highlighted in orange.

	TG1	TG2	TG3	TG4
Q1	7.73E-01	3.20E-01	3.48E-01	7.49E-01
Q2	1.08E-01	2.03E-02	1.96E-01	4.90E-01
Q3	4.43E-01	7.39E-02	8.84E-02	5.59E-01
Q4	2.77E-01	4.98E-02	3.04E-01	9.50E-01
Q5	2.52E-03	2.97E-03	3.64E-01	7.94E-01
Q6	1.75E-02	3.38E-02	7.71E-01	4.00E-01
Q7	1.03E-01	1.80E-03	2.59E-01	1.03E-01
Q8	2.74E-02	1.50E-04	3.11E-03	5.34E-01
Q9	4.07E-05	2.87E-02	7.63E-03	9.88E-01
Q10	4.42E-07	4.95E-05	9.55E-02	6.39E-01
Q11	3.52E-02	1.17E-01	2.55E-02	3.21E-01
Q12	1.06E-02	1.97E-02	9.84E-01	3.03E-01
Q13	1.19E-06	5.31E-08	8.38E-07	6.48E-01
Q14	4.22E-03	9.81E-04	1.69E-01	9.36E-01
Q15	6.49E-03	1.94E-03	4.21E-02	3.91E-01
Q16	6.52E-05	3.77E-07	2.07E-01	9.79E-01
Q17	1.28E-01	1.42E-01	8.03E-03	5.16E-02
Q18	4.48E-02	8.15E-03	1.87E-01	2.55E-02
Q19	5.58E-05	5.90E-05	2.60E-02	7.70E-01
Q20	5.24E-05	8.03E-05	1.67E-03	1.95E-01

11) Difference between week 1 and week 2 item characteristic curves

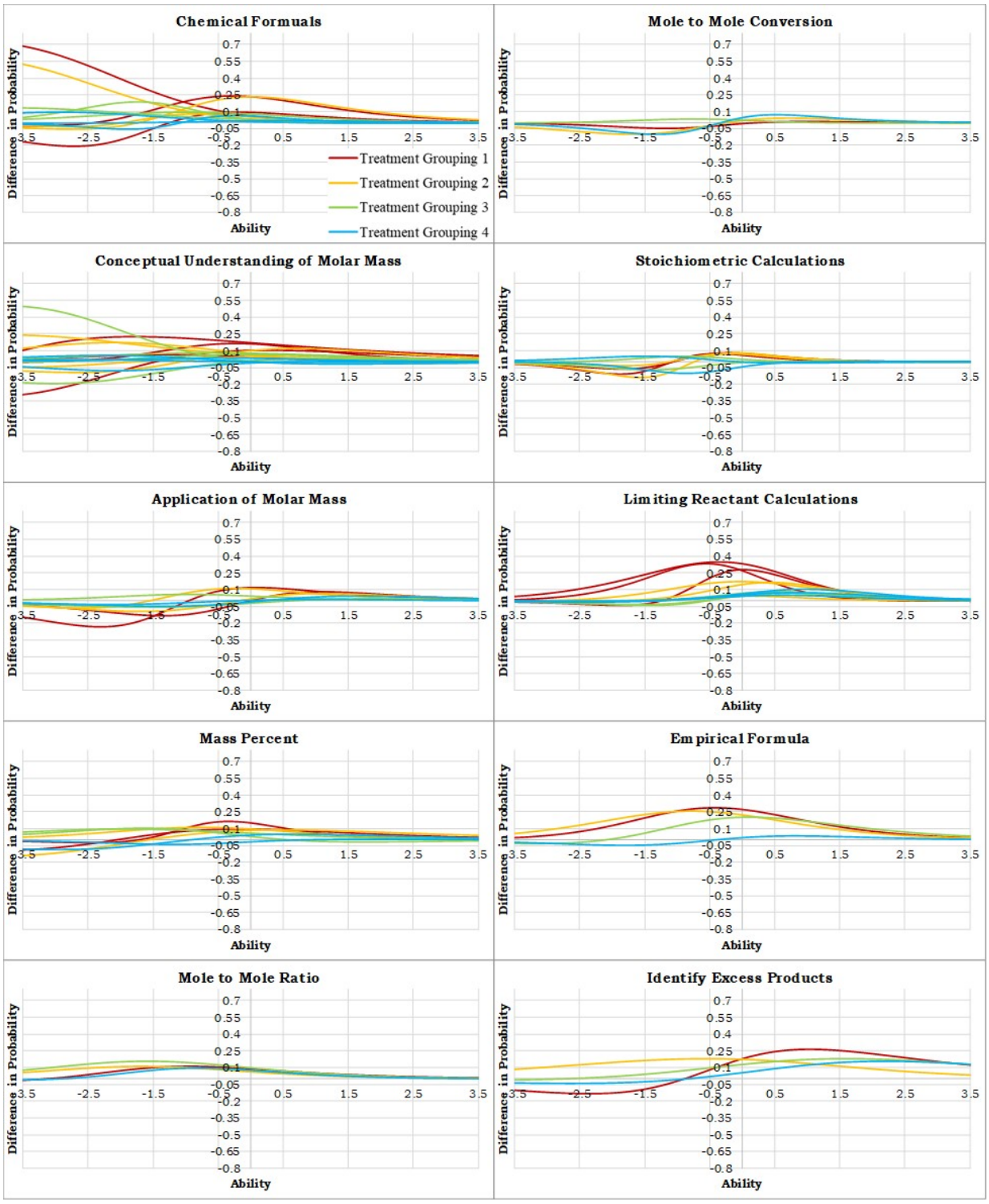


Fig. 11 Difference between week 1 and week 2 ICC's plotted within content areas.

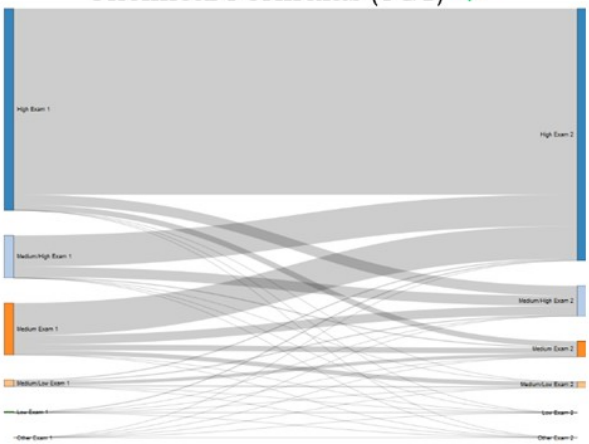
12) Multimode analysis

In addition to the quantitative methods used to assess each treatment, qualitative measures were also used to access student growth. Multimode grading was previously conducted on this exam, and the results of this grading scheme were also applied to the research questions here-in (Murphy et al.). As a brief summary, multimode grading was conducted in four key steps labelled [1]-[4]. [1] First, the response options (A-D) of each item were analysed and the ability level of a student who would choose that response was ordinally estimated from the following options: high, medium/high, medium, medium/low, or low. [2] The exam questions were then ordered based on content progression. Content progression was not necessarily correlated with item difficulty, rather early content questions only required foundational knowledge where later content questions required an understanding of the earlier foundational knowledge to answer correctly. [3] Then, aided by the ordering of questions based on content progression, questions were grouped into broader content areas. [4] From there, within each content area, student ability was again estimated for each content topic based on possible response patterns. This method was specifically used to assess changes in student score within specific content areas.

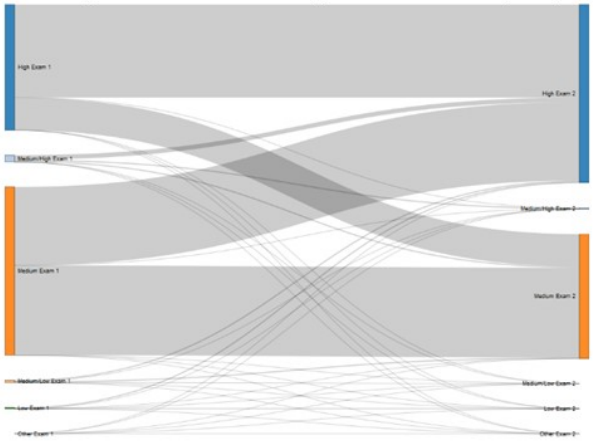
12.1) Content-specific multimode analysis

Students' ability within each content area each week was determined using the multimode method (Murphy et al.). After content-specific ability levels were determined, Sankey diagrams were constructed to visualize ability migration from week 1 (left column) to week 2 (right column). The height of each ability level (High, Medium/High, Medium, Medium/Low, Low, Other) corresponds to the population of that ability level. The thickness of the grey connections between week 1 and week 2 reference the number of students who made that specific migration. These shifts are shown in the figure below for each treatment grouping and most content areas. Two content areas ("Empirical Formula" and "Identify Excess Products") are not included as the multimode analysis was not able to assign an ability estimate for those content areas (Murphy et al.). The checkmark (✓) and cross (✗) on the top of each image reflect whether overall improvement was seen for the diagram.

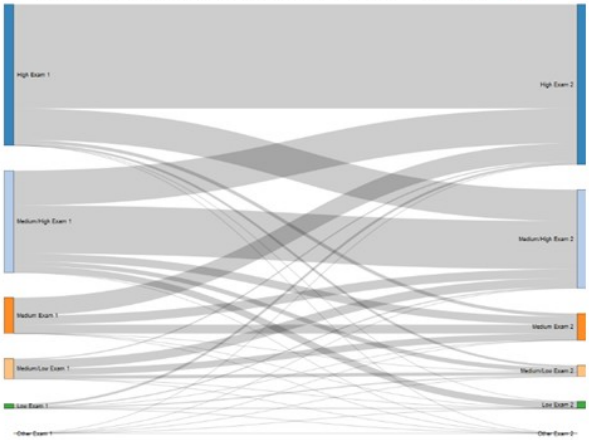
Chemical Formulas (TG1) ✔



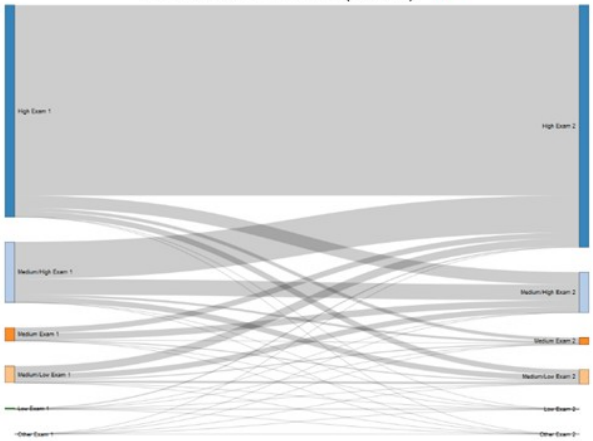
Conceptual Understanding of Molar Mass (TG1) ✘



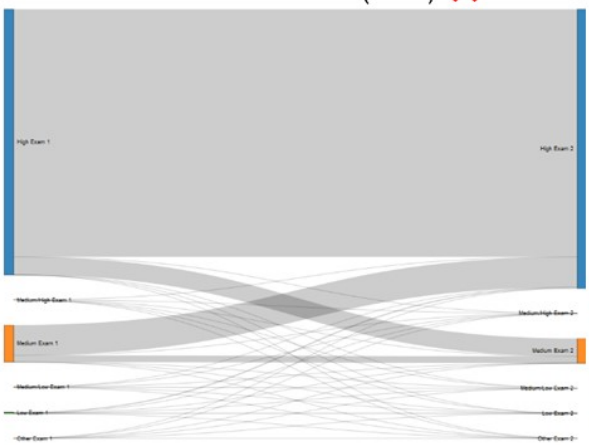
Application of Molar Mass (TG1) ✘



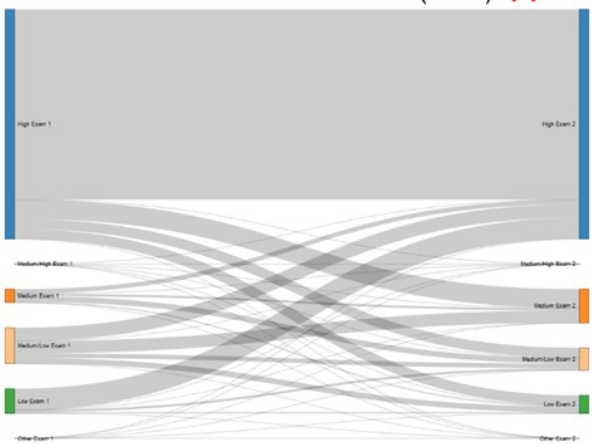
Mass Percent (TG1) ✔



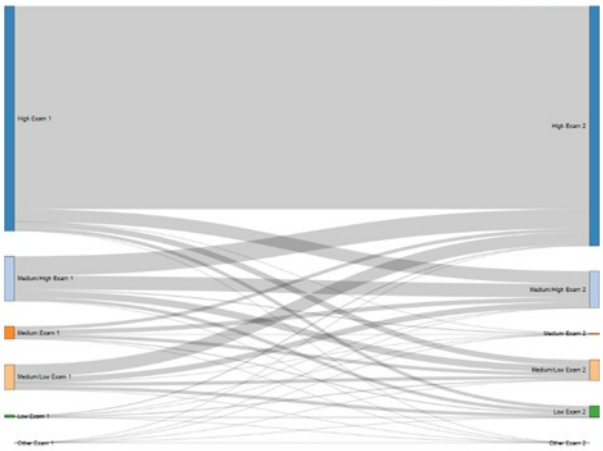
Mole to Mole Ratio (TG1) ✘



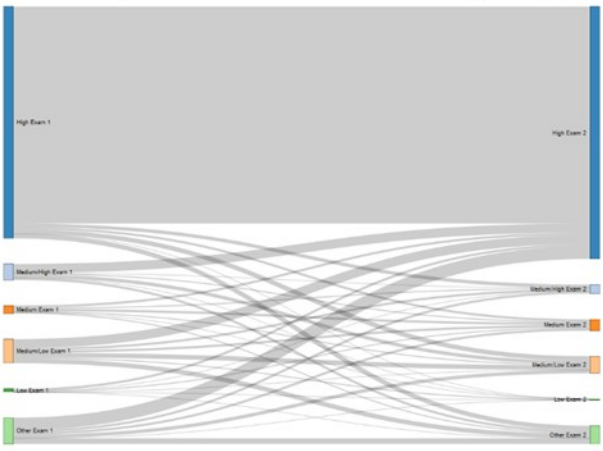
Mole to Mole Conversion (TG1) ✘



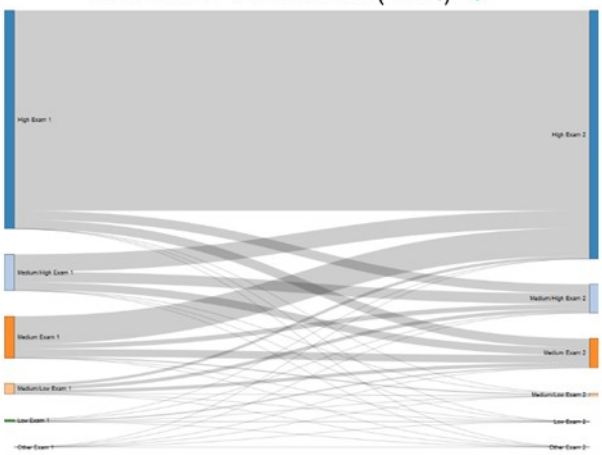
Stoichiometric Calculations (TG1) ✗



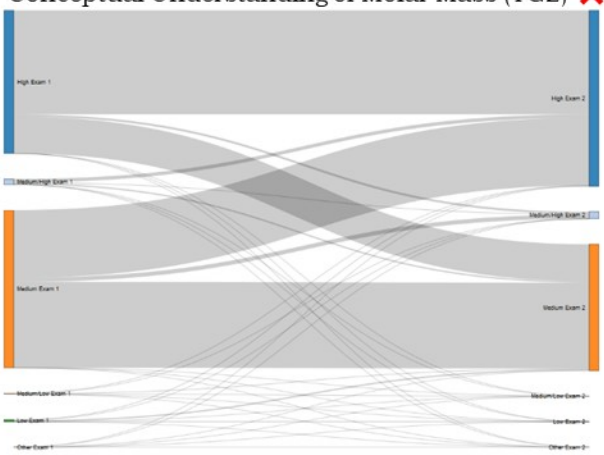
Limiting Reactant Calculations (TG1) ✗



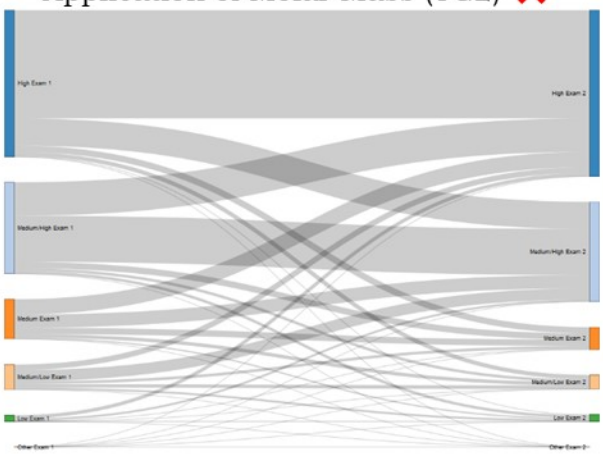
Chemical Formulas (TG2) ✓



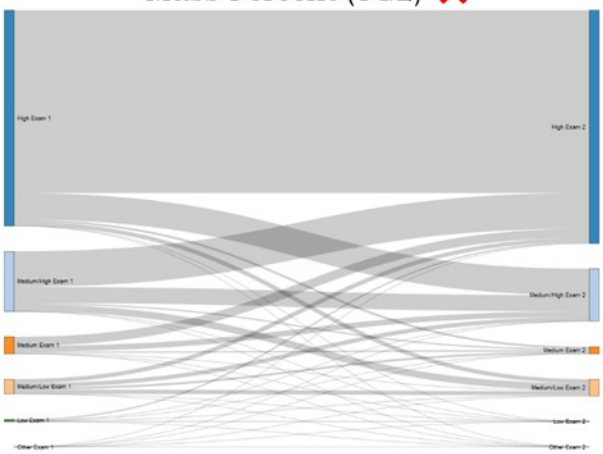
Conceptual Understanding of Molar Mass (TG2) ✗



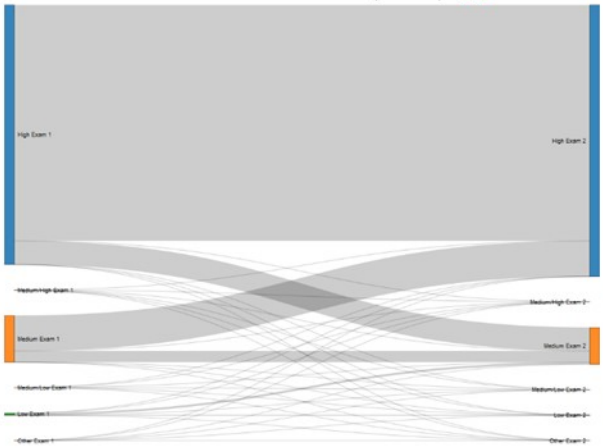
Application of Molar Mass (TG2) ✗



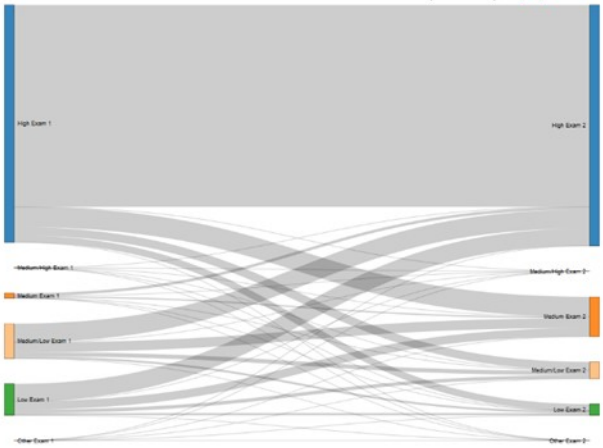
Mass Percent (TG2) ✗



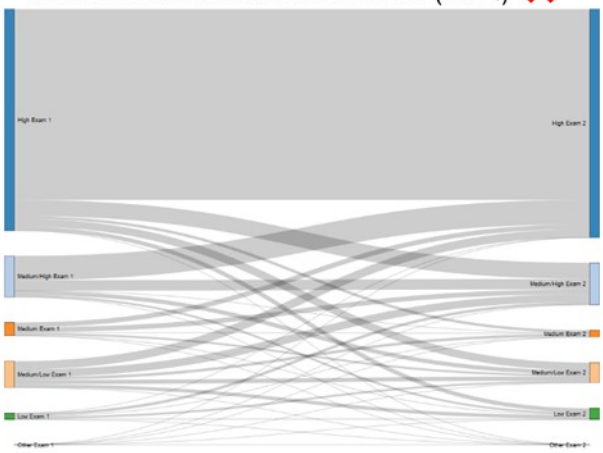
Mole to Mole Ratio (TG2) ❌



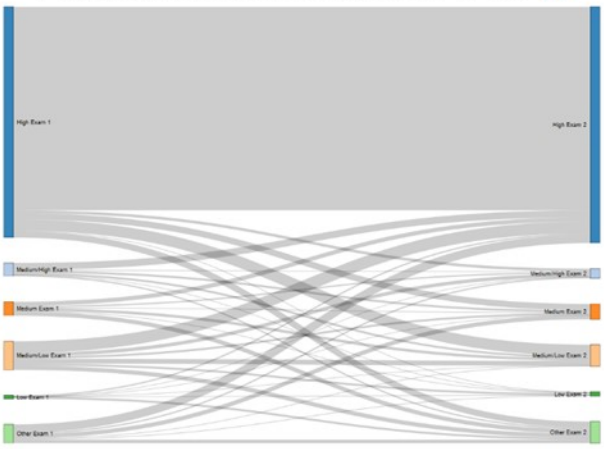
Mole to Mole Conversion (TG2) ❌



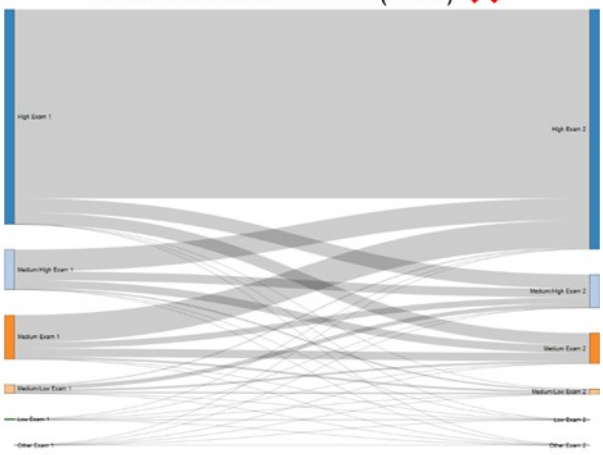
Stoichiometric Calculations (TG2) ❌



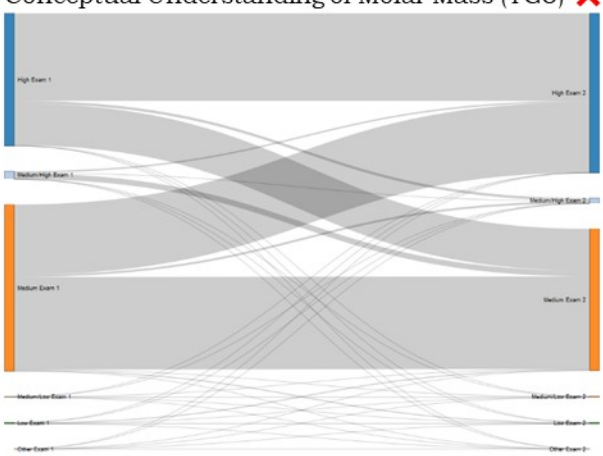
Limiting Reactant Calculations (TG2) ❌



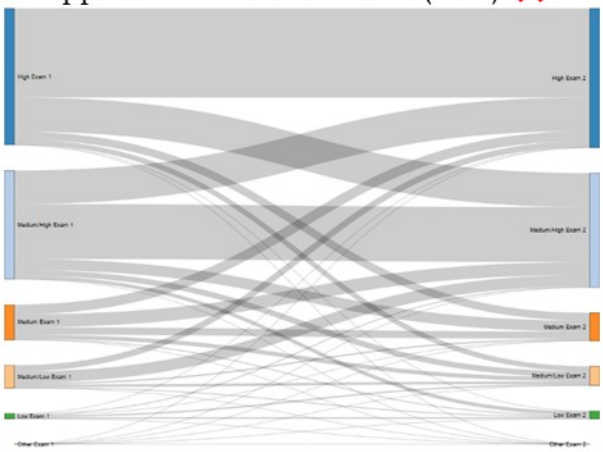
Chemical Formulas (TG3) ❌



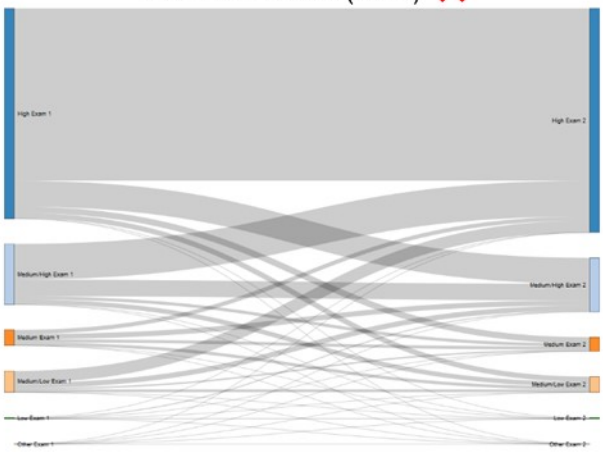
Conceptual Understanding of Molar Mass (TG3) ❌



Application of Molar Mass (TG3) ❌

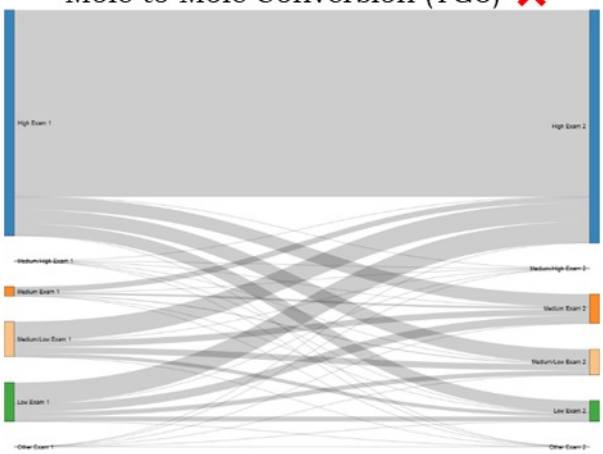


Mass Percent (TG3) ❌

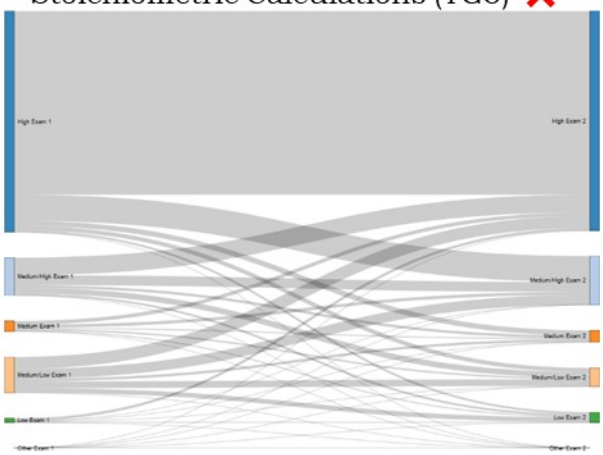


Mole to Mole Ratio (TG3) ❌

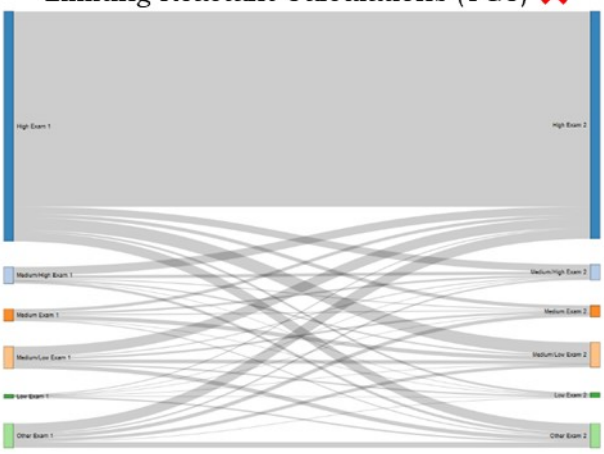
Mole to Mole Conversion (TG3) ❌



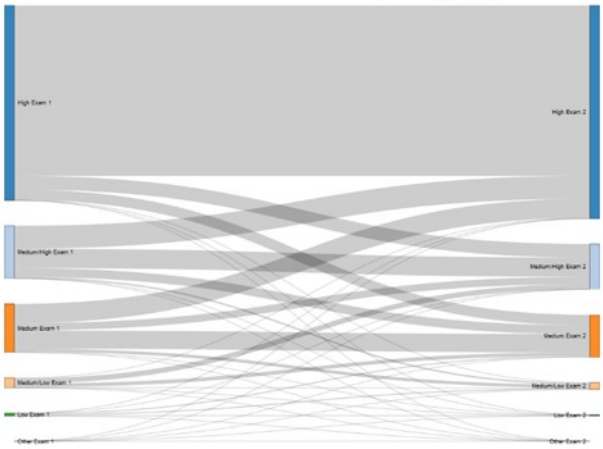
Stoichiometric Calculations (TG3) ❌



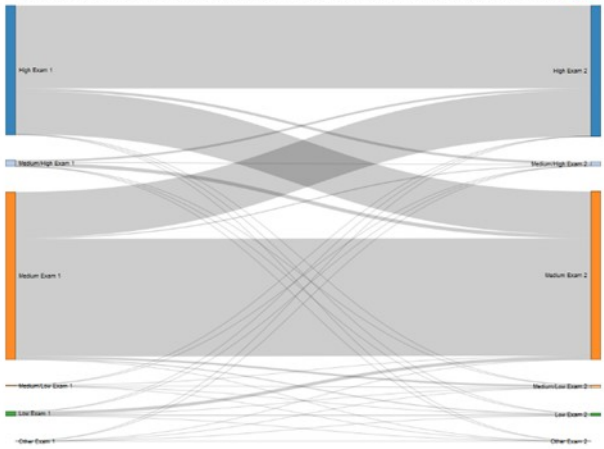
Limiting Reactant Calculations (TG3) ❌



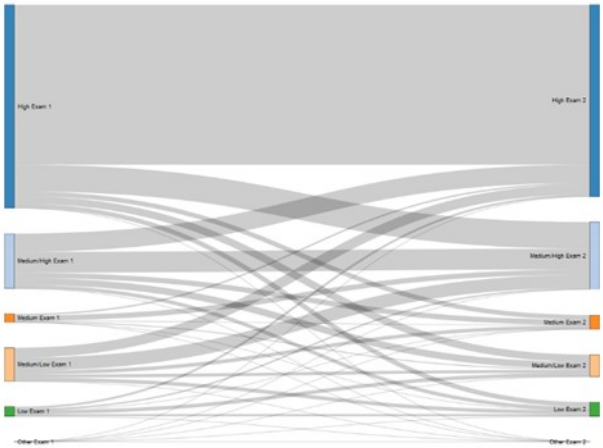
Chemical Formulas (TG4) ✗



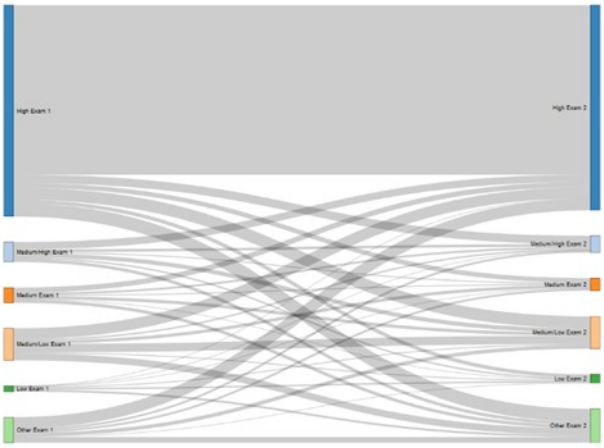
Conceptual Understanding of Molar Mass (TG4) ✗



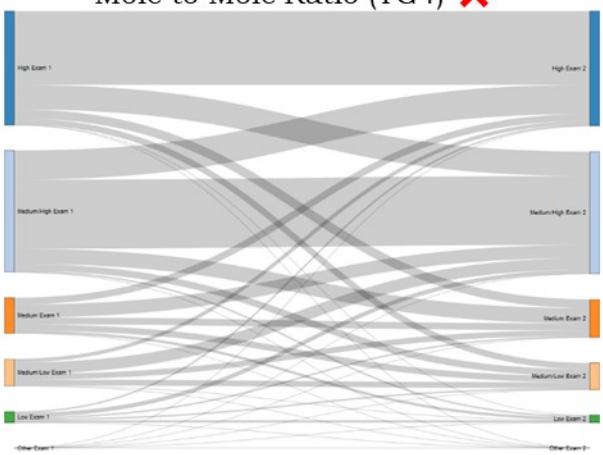
Application of Molar Mass (TG4) ✗



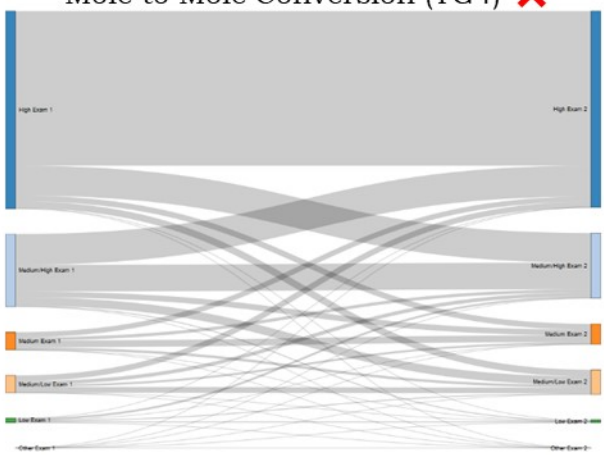
Mass Percent (TG4) ✗



Mole to Mole Ratio (TG4) ✗



Mole to Mole Conversion (TG4) ✗



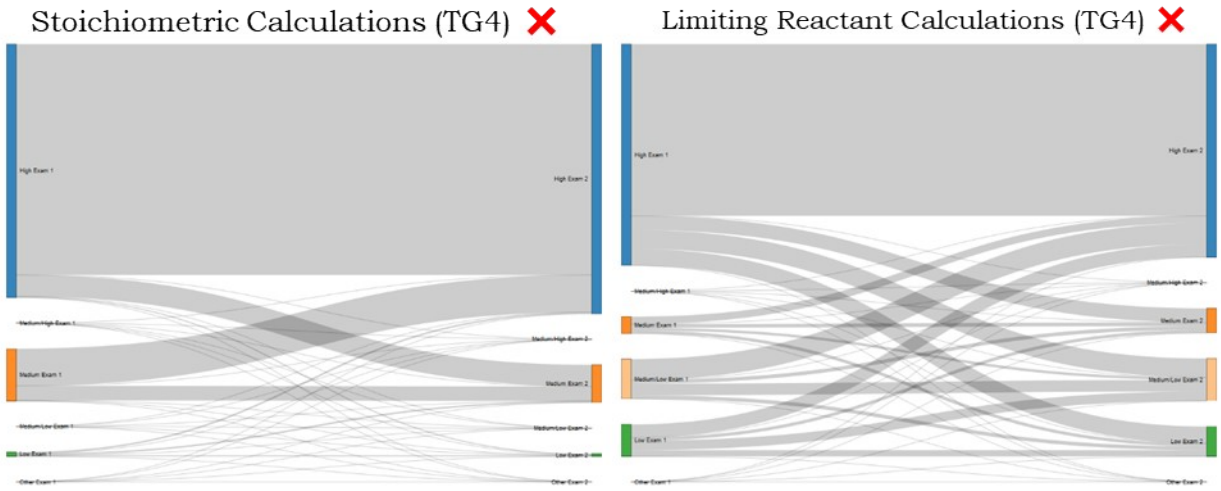


Fig. 12 Multimode ability and migration between weeks for each content area and Treatment Grouping (TG). Diagrams where dramatically more improvement was seen are marked with a "✓" where diagrams where improvement was canceled out by decline is marked with a "✗".

13) ESI References

An X. and Yung Y., (2014), Item Response Theory : What It Is and How You Can Use the IRT Procedure to Apply It. SAS Institute Inc., 1–14.

Bock R. D., (2005), A Brief History of Item Response Theory. Educational Measurement: Issues and Practice, 16(4), 21–33, DOI: 10.1111/j.1745-3992.1997.tb00605.x.

Cooper M. M., Cox C. T., Nammouz M., Case E., and Stevens R., (2008), An assessment of the effect of collaborative groups on students' problem-solving strategies and abilities. J Chem Educ, 85(6), 866–872, DOI: 10.1021/ed085p866.

Cornelius A., Brewer B., and Raalte J., (2007), Applications of multilevel modeling in sport injury rehabilitation research. Int J Sport Exerc Psychol, 5(4), 387–405, DOI: 10.1080/1612197x.2007.9671843.

Doran H. C. and Lockwood J. R., (2006), Fitting Value-Added Models in R. Journal of Educational and Behavioral Statistics, 31(2), 205–230, DOI: 10.3102/10769986031002205.

Glynn S. M., (2012), International assessment: A Rasch model and teachers' evaluation of TIMSS science achievement items. J Res Sci Teach, 49(10), 1321–1344, DOI: 10.1002/tea.21059.

Hambleton R., Rogers H., and Swaminathan H., (2012), Fundamentals of Item Response Theory, Sage Publications.

Holland P. W. and Wainer H., (2009), Differential item functioning, Routledge.

Holme T. and Murphy K., (2011), Assessing Conceptual and Algorithmic Knowledge in General Chemistry with ACS Exams. J Chem Educ, 88(9), 1217–1222, DOI: 10.1021/ed100106k.

IBM Corp, (2017), IBM SPSS Statistics for Windows.

Kendhammer L., Holme T., and Murphy K., (2013), Identifying differential performance in general chemistry: Differential item functioning analysis of acs general chemistry trial tests. J Chem Educ, 90(7), 846–853, DOI: 10.1021/ed4000298.

Kendhammer L. K. and Murphy K. L., (2014), Innovative Uses of Assessments for Teaching and Research, American Chemical Society, pp. 1–4, DOI: 10.1021/bk-2014-1182.ch001.

Laursen S. L. and Weston T. J., (2014), Trends in Ph.D. productivity and diversity in top-50 U.S. chemistry departments: An institutional analysis. J Chem Educ, 91(11), 1762–1776, DOI: 10.1021/ed4006997.

Lee S. and Suh Y., (2018), Lord's Wald Test for Detecting DIF in Multidimensional IRT Models: A Comparison of Two Estimation Approaches. J Educ Meas, 55(2), 328–353, DOI: 10.1111/jedm.12178.

Lord F. M., (1980), Applications of item response to theory to practical testing, Erlbaum Associates.

Murphy K., Schreurs D., Teichert M., Luxford C., and Schneider J., A Comparison of Observed Scores, Partial Credit Schemes, and Modeled Scores Among Chemistry Students of Different Ability Groupings. Manuscript in preparation

O'Connell A. A. and McCoach D. B., (2004), Applications of hierarchical linear models for evaluations of health interventions: Demystifying the Methods and Interpretations of Multilevel Models. Eval Health Prof, 27(2), 119–151, DOI: 10.1177/0163278704264049.

Singer J. D., (1998), Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models. Journal of Educational and Behavioral Statistics, 23(4), 323, DOI: 10.2307/1165280.

Weaver G. C. and Sturtevant H. G., (2015), Design, Implementation, and Evaluation of a Flipped Format General Chemistry Course. J Chem Educ, 92(9), 1437–1448, DOI: 10.1021/acs.jchemed.5b00316.

Zumbo B., (1999), A handbook on the theory and methods of differential item functioning (DIF).