Optimizing testing feedback in introductory chemistry: a multi-treatment study exploring varying levels of assessment feedback and subsequent performance

Journal:	*Chemistry Education Research and Practice*
Manuscript ID	RP-ART-03-2024-000077.R1
Article Type:	Paper
Date Submitted by the Author:	09-May-2024
Complete List of Authors:	Murphy, Kristen; University of Wisconsin-Milwaukee, Chemistry and Biochemistry Schreurs, David; ACS Exams, Teichert, Melonie; United States Naval Academy, Chemistry Luxford, Cynthia; Texas State University San Marcos, Department of Chemistry and Biochemistry Trate, Jaclyn; University of Wisconsin-Madison, Department of Chemistry Harshman, Jordan; Auburn University, Chemistry and Biochemistry Schneider, Jamie; University of Wisconsin-River Falls, Chemistry

ARTICLE

Optimizing testing feedback in introductory chemistry: a multi-treatment study exploring varying levels of assessment feedback and subsequent performance

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Kristen L. Murphy,[a] David G. Schreurs,[a] Melonie A. Teichert,[b] Cynthia J. Luxford,[c] Jaclyn M. Trate,[‡a] Jordan T. Harshmann[§d] and Jamie L. Schneider*[d]

Providing students with feedback on their performance is a critical part of enhancing student learning in chemistry and is often integrated into homework assignments, quizzes, and exams. However, not all feedback is created equal, and the type of feedback the student receives can dramatically alter the utility of the feedback to reinforce correct processes and assist in correcting incorrect processes. This work seeks to establish a ranking of how eleven different types of testing feedback affected student retention or growth in performance on multiple-choice general chemistry questions. These feedback methods ranged from simple noncorrective feedback to more complex and engaging elaborative feedback. A test-retest model was used with a one-week gap between the initial test and following test in general chemistry I. Data collection took place at multiple institutions over multiple years. Data analysis used four distinct grading schemes to estimate student performance. These grading schemes included dichotomous scoring, two polytomous scoring techniques, and the use of item response theory to estimate students' true score. Data were modeled using hierarchical linear modeling which was set up to control for any differences in initial abilities and to determine the growth in performance associated with each treatment. Results indicated that when delayed elaborative feedback was paired with students being asked to recall/rework the problem, the largest student growth was observed. To dive deeper into student growth, both the differences in specific content-area improvement and the ability levels of students who improved the most were analyzed.

Introduction and theory

General chemistry is often taught in large lecture sections influencing instructor choices to utilize multiple-choice exams, frequently with limited feedback. Feedback provided after a multiple-choice exam is often a passive activity whereby faculty post scores, an answer key, and/or worked out solutions, but it is up to the individual student whether and how to engage with that feedback. Thus, understanding the best ways to use and provide feedback for multiple-choice exams could greatly help instructors employ best practices to maximize student learning through testing.

Testing Effect is commonly cited as a robust concept established through numerous research studies (Karpicke and Roediger, 2008; Rowland, 2014, Todd et al., 2021). Testing Effect is the finding that retrieval of information (testing) leads to better retention compared to restudying of the material, which suggests that testing could also be considered a learning tool in addition to an assessment tool (Karpicke, 2012; American Psychological Association (APA), 2021). Unlike the content often tested in a general chemistry course, much of the Testing Effect literature involves materials that require fact-based retrieval like word lists, symbol-word pairs, and reading comprehension of a prose passage (Wheeler et al., 2003; Roediger and Karpicke, 2006; Coppens et al., 2011). Although Karpicke and Aue argued that Testing Effect is beneficial for more complex materials citing several examples, Van Gog and Sweller challenged this notion suggesting that there was a limit to testing benefits with complex problem solving that requires more than recall but also generation and reconstruction of information (Karpicke and Aue, 2015; Van Gog and Sweller, 2015; Van Gog et al., 2015). Van Gog *et al.* proposed that the drop off in efficacy of Testing Effect was because earlier fact-based studies required declarative memory encoding, whereas problem solving required declarative schema construction and inclusion of procedural memory. In three experiments, Van Gog et al. had university students from various programs participate in learning about and solving problems related to circuits. In the fourth experiment, students in a math course were recruited to solve probability calculations. The goal of these experiments was to assess restudy of problems versus testing with the problems. In each experiment, ANOVA analysis of a final test one week after the initial treatment revealed no significant differences between the populations that restudied the material versus the populations that took a test of the material. Although this study seemed to contradict the many Testing

[a] University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53211
[b] United States Naval Academy, Annapolis, Maryland 21402
[c] Texas State University, San Marcos, Texas 78666
[d] University of Wisconsin-River Falls, River Falls, Wisconsin 54022

Effect studies, there were several limitations to this study. Limitations included relatively short acquisition times for novice students to learn about circuits, small sample sizes with each experiment consisting of 57-120 students split between multiple treatment conditions, and final judgement based on performance of only a few items. They do suggest that further research with the addition of corrective feedback and with students with stronger prior knowledge of the material would be beneficial to elucidate the boundaries of Testing Effect with problem solving.

Testing with feedback has been shown to enhance Testing Effects. One issue with multiple-choice testing, and testing in general, is that choosing an answer (even a wrong answer) creates a relatively strong memory when completed under testing conditions. Without corrective feedback, incorrect answers are likely to show up again (Roediger and Marsh, 2005). Butler and Woodward provide a nice summary of the literature using task-level feedback to promote learning of facts and concepts (Butler and Woodward, 2018), defining task-level feedback as "feedback provided by an external agent after individual events that require memory retrieval". They also provide a careful analysis and theoretical interpretation of testing feedback literature that contains a variety of feedback types and timing, each factor potentially changing the efficacy of outcomes. Our research group also provided a summary of testing feedback literature in an ACS Symposium chapter on testing (Schneider et al., 2014). To summarize the testing feedback literature: delayed corrective feedback produces better retention compared to immediate feedback, but immediate feedback may be useful with initial acquisition of information, especially for poor performing learners (Wright and Gescheider, 1970; Kulhavy and Anderson, 1972; Butler et al., 2007; Butler and Roediger, 2008; Mullet et al., 2014). For high efficacy of testing feedback on future performance, corrective feedback type should contain the original item, student's answer, and the correct answer (Hintzman, 2010). The feedback type with the most discrepant research was feedback that provided elaboration or explanation for the answer choices, with some research showing no improvement with elaboration while others showing moderate improvement for certain types of questions (Moreno, 2004; Butler et al., 2013). For future research, Butler and Woodward suggest the following: further investigation of errors and specifically feedback effects on different types of errors, investigating strategies that require generation and active processing during feedback, shifting from verbatim repeat testing to more transfer situations, and assessing the influence of metacognitive cues on confidence judgements and processing of feedback (Butler and Woodward, 2018). With all of these studies, there is a hope that this research will continue to inform educational practices, particularly in the context of science instruction (Risley, 2007; Henderson and Harper, 2009; Knaus et al., 2009; Andaya et al., 2017).

Our team endeavoured to use this rich literature to establish a research program investigating testing feedback efficacy on future performance outcomes in introductory chemistry courses. As practitioners and researchers, we wanted to design our study in a way that would directly link to classroom testing/practice testing applications in most introductory science and math courses and would apply robust, reproducible methods to add to the testing feedback literature. To establish experimental conditions with student prior knowledge and motivation similar to a multiple-choice testing scenario, we elected to use a practice test given to students enrolled in general chemistry courses at multiple institutions a few weeks prior to the course final exam. After and/or during the practice test, students were given some feedback condition. About 1 week later, the same students took an isomorphic practice test. The change in test performance, including the change in test performance by different groups of students, was then studied using a variety of modelling techniques to investigate the efficacy of different feedback mechanisms.

Research questions

Methods and analysis for this study were structured to answer the following research questions:

(1) How do different types of feedback relate to student growth in exam performance, and which type of feedback is most effective?

(2) How does the growth related to each type of feedback differentially benefit students of particular ability levels?

(3) Does the content area of the question influence student growth, and how is that different between the treatments?

Methodology

Volunteer students enrolled in general chemistry at five institutions (labelled I1-I5) over 8 years participated in the test-retest study. A brief summary of each of these institutions is included in the Electronic Supplementary Information (ESI section 1). Unique student groups were assigned to each feedback condition, but within the feedback condition the same group of students took the initial and final practice tests (between-within design). Two isomorphic exams consisting of introductory chemistry concepts [chemical composition (6 items); gram, mole, molecule conversions (6 items); reaction stoichiometry (4 items); and limiting reactant stoichiometry (4 items)] were used for data collection (Murphy et al.). Items were written to assess both quantitative and conceptual reasoning, including interpretation of particulate-level drawings. Previous work has provided evidence of high validity for the data produced by these items when administered to general chemistry students (Murphy et al.; Schreurs et al., 2024; Trate et al., 2020). Reliability was also assessed through Cronbach's alphas and average inter-item correlations which are included in ESI section 4.2. Practice Tests are available upon request from the corresponding author. Detailed information on the scoring of these tests is detailed elsewhere and in ESI section 12 (Murphy et al.). Student consent was obtained per IRB at each institution, and all students who did not provide consent were excluded from analysis.

General practice methodology: Delayed Noncorrective (DNC)

The general data collection strategies were held constant at each testing site with instructors at each institution proctoring the practice exams during out of class sessions and/or laboratory sessions. Faculty proctors were given recruitment information and PowerPoint presentations to guide their introduction of the testing protocols. Students were invited to participate in the study about 3-4 weeks prior to the course final exam to help prepare for the final exam in some key course concepts. Students had up to 60 minutes to complete the initial 20-item multiple-choice Exam A. The proctors scanned the exam scoring sheets and sent them to the research team to score using electronic scoring software (Gravic, 2023). A score feedback sheet sorted by pre-assigned research student IDs was generated by the research team and posted for the participating students on a course management site (like D2L or Canvas) within 2-3 days of initial testing by the proctor. The feedback sheet indicated how many items were correct out of 20 and how many items were correct in each of the four topic areas. One week after taking the initial test, students returned to take an isomorphic 20-item multiple-choice Exam B with the same 60-minute limitation. Exam B had the same order of items with the same stem language but with different elements, compounds, and/or quantities. The answer choices also had the same language and same processing errors; however, the order of the answer choices was different to encourage re-processing of items rather than recalling answer choice order. Exam B was processed like Exam A with a score feedback sheet posted to the course management site within 2-3 days of testing. This general practice testing methodology constituted what we describe as Delayed Noncorrective feedback since students received their scores after taking the exam, but they did not receive information on which questions they missed, what their errors were, or the correct answers. Students were given access to overall score and subscores for all treatments in this study. This experimental condition is Treatment 1.

Validation steps were taken to ensure the week 2 exam (Exam B) was comparable to the week 1 exam (Exam A). The student sample analysed for the validation included 2,025 students who completed Exam A and 219 students who completed Exam B during week 1, without immediate feedback. The exams were graded dichotomously (correct or incorrect) and polytomously. Two partial-credit polytomous grading schemes were used and were dubbed “open” and “hierarchy”. Partial credit values were determined by expert raters. The open partial credit values could be any permutation of 0, 0.25, 0.5, or 1 where values could be used multiple times or excluded entirely. Questions under the hierarchy scheme required each of the values (0, 0.25, 0.5, 1) to be used once per question for the 4 answer choices. More details about these partial credit methods are provided elsewhere (Murphy et al.). To ensure cloning, test performance was compared between both versions and is included in ESI section 4.2. All analyses pointed towards both exams performing comparably.

Corrective feedback methodology

Each corrective feedback condition was added to the general practice testing methodology (Treatment 1) with changes occurring either during week 1 testing (immediate feedback conditions) or between week 1 and 2 testing (delayed feedback conditions). All Treatments utilized the same exams: A then B. The delayed noncorrective feedback condition described by the general methodology (Treatment 1) served as the control (e.g., item correctness was not indicated). In all, there were 10 different corrective feedback conditions (Treatments 2-11) described in Table 1. Only students who fully answered all 20 items on both week 1 and week 2 exams and who fully participated in the corrective feedback were included in the analysis.

Immediate Feedback Assessment Technique (IFAT)

During Treatments 2, 6, and 7 week 1 testing, students were instructed to bubble in their first answer choice on an electronically scored answer sheet and then to use a commercially available Immediate Feedback Assessment Technique (IFAT) form to scratch off a waxy coating to reveal the correctness of their answer choice (Epstein Educational Enterprises). The intent of the IFAT form is to provide immediate feedback to students so that they may correct their process before progressing to the next item. IFAT was the only corrective feedback mechanism employed for Treatment 2. The specific directions provided to students for using the IFAT forms are provided in ESI section 2.

Delayed Corrective (DC)

In Treatments 3-11, some form of delayed corrective (DC) feedback was provided 2-6 days after week 1 testing but before week 2 testing. Electronic platforms (Web Assign and Qualtrics) were used to deliver the delayed feedback conditions. In the delayed feedback conditions, students were shown each item one at a time, and students either entered their original Exam A answer choice (student prior answer choices were provided to them by the researchers) or the students were asked to recall/rework the problem and enter their answer choice. Based on testing feedback research optimal conditions, students would view the original item, their answer, the correctness of their answer, and the correct answer during the testing feedback process (Butler and Woodward, 2018). In Treatments 3, 5, 7, 9, and 11, student answer choices for week 1 were provided to the students’ course management site for students to input into the online feedback test to get item by item corrective feedback (Answers Provided, AP). In Treatments 4, 6, 8 and 10, students Recalled/Reworked (RR) each item during the Exam A feedback. In the delayed corrective condition (Treatment 3), students were given the following types of messages based on the correctness of their answer:

> *“Your answer choice was incorrect. The correct answer choice was D 55.3%.”*

"Your answer choice of D 55.3% was correct."

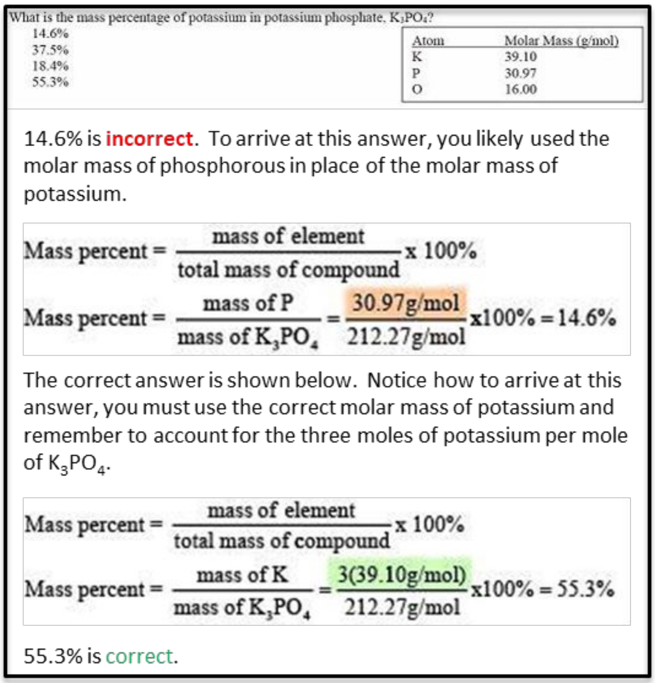
Answer Until Correct (AUC)

In the delayed Answer Until Correct (AUC) condition (Treatments 4-7), students were instructed to continue to answer each item one at a time until they chose the correct answer. After each attempt, students were given the following messages based on the correctness of their answer:

> *"You answered A. This is incorrect. Please try another answer."*
>
> *"You answered B. This is correct. Good job! Please go on to the next problem."*

Elaborative Feedback (EF)

In the Elaborative Feedback (EF) conditions, students were given additional information on the likely mistake they made based on the answer choice selected and a possible solution to obtain the correct answer. In Treatments 8 & 9 (EF1), this took the form of only one answer being elaborated to the students. The elaborative feedback was developed from response process research done with student interviews (Schreurs et al., 2024; Trate et al., 2020). An example of this is provided below and an additional example for a conceptual item is included in ESI section 6:



What is the mass percentage of potassium in potassium phosphate, K_3PO_4?

14.6%
37.5%
18.4%
55.3%

Atom	Molar Mass (g/mol)
K	39.10
P	30.97
O	16.00

14.6% is **incorrect**. To arrive at this answer, you likely used the molar mass of phosphorous in place of the molar mass of potassium.

$$\text{Mass percent} = \frac{\text{mass of element}}{\text{total mass of compound}} \text{x } 100\%$$

$$\text{Mass percent} = \frac{\text{mass of P}}{\text{mass of } K_3PO_4} = \frac{30.97\text{g/mol}}{212.27\text{g/mol}} \text{x}100\% = 14.6\%$$

The correct answer is shown below. Notice how to arrive at this answer, you must use the correct molar mass of potassium and remember to account for the three moles of potassium per mole of K_3PO_4.

$$\text{Mass percent} = \frac{\text{mass of element}}{\text{total mass of compound}} \text{x } 100\%$$

$$\text{Mass percent} = \frac{\text{mass of K}}{\text{mass of } K_3PO_4} = \frac{3(39.10\text{g/mol})}{212.27\text{g/mol}} \text{x}100\% = 55.3\%$$

55.3% is correct.

In Treatments 10 & 11 (EF2), students were given the same elaborative feedback information if they were incorrect on the first answer choice but instead of the correct answer solution they were told to "click on the next page to try this question again". On the second attempt, they got the same feedback as the students did on the elaborative feedback condition (1 answer).

All 11 treatment conditions are summarized in Table 1, which is ordered to approximately mimic an increasing amount of feedback to the student (*e.g.* Treatment 11 is substantially more feedback to the student than Treatment 1, for example).

Statistical analysis of exam score data

The classical treatment of test-retest data comparing different feedback mechanisms would be to utilize a repeated measures ANOVA analysis. Because of the number of treatments (11 Treatment conditions) and the partial-credit scoring options, a more sophisticated approach was taken to more carefully discern differences in student performance. Specifically, our analysis of the exam data necessitated the use of many mathematical tools such as Hierarchal Linear Modelling (HLM) and Item Response Theory (IRT). HLMs were chosen as the primary modelling technique because of their efficiency when investigating the relationships within and between hierarchical levels such as students within treatments over time (Woltman et al., 2012). For some analyses, HLM was paired with IRT. IRT attempts to transform the observed performance on a given assessment into a more accurate prediction of underlying ability. A commonly used statistic within IRT is the Lord's Wald test which provides a metric for showing significant difference between two groups on a common item (e.g., do week 1 and week 2 perform the same on this question?). Differences between two groups can also be visualized using Item Characteristic Curves (ICCs). ICCs plot student's ability levels against the probability of answering each item correctly (Schurmeier et al., 2010). If the treatment improved student performance, that ICC would be expected to appear higher (more likely to answer correctly) at most ability levels. Effect size for each ability level can be calculated using Cohen's h to compare the proportion of students at each ability level who are likely to answer correctly (based on the ICC) before and after the treatment (Cohen, 1988). Further background on HLM, IRT, and the Lord's Wald test are included in ESI section 3.1-3.2.

Table 1 Brief description of treatments and the indexing used for all future analysis.

Code	Treatment*	Description
1	DNC Only	Delayed Noncorrective (DNC)
2	IFAT	IFAT and DNC
3	DC-AP	Delayed Corrective no Elaborative Feedback (prior answer choices provided) and DNC
4	AUC-RR	Delayed AUC no Elaborative Feedback (asked to rework/recall) and DNC
5	AUC-AP	Delayed AUC no Elaborative Feedback (prior answer choices provided) and DNC
6	IFAT + AUC-RR	IFAT and Delayed AUC no Elaborative Feedback (asked to rework/recall) and DNC
7	IFAT + AUC-AP	IFAT and Delayed AUC no Elaborative Feedback (prior answer choices provided) and DNC
8	EF1-RR	Delayed Elaborative Feedback (1 answer) (asked to rework/recall) and DNC
9	EF1-AP	Delayed Elaborative Feedback (1 answer) (prior answer choices provided) and DNC
10	EF2-RR	Delayed Elaborative Feedback (2 answer) (asked to rework/recall) and DNC
11	EF2-AP	Delayed Elaborative Feedback (2 answer) (prior answer choices provided) and DNC

*all treatments involved DNC

Software specifications

The analyses required the implementation of IRT and the construction of HLMs. All statistical analysis was conducted using R 3.6.2 (R Core Team, 2022). To import raw data from excel, the R package 'readxl' was used (Wickham and Bryan, 2019). After all analyses, the R package 'xlsx' was used to export results from R back into excel (Dragulescu and Arendt, 2020). The 'ltm' package was used to construct the IRT models and 'difR' was used to compare the models using Lord's Wald test (Rizopoulos, 2006; Magis et al., 2010). HLMs were created using both the 'nlme' and 'lme4' package (Bates et al., 2015; Pinheiro et al., 2020). After the HLMs were formed, the 'arm' package was used to estimate the standard error associated with the random effects (Gelman and Su, 2020).

Pilot hierarchical linear models

To determine the optimal model to fit the data, four pilot models were constructed and compared. The initial model was the simplest and was tested against 3 iterations of building up the model through the addition of fixed and random effects. The process of these models is explained in detail in ESI section 3.3 but the optimal model (referred to as "m2") used: student initial performance, student improvement between weeks, and a random effect based on the student's treatment.

Results

The average exam scores and total number of students who took the exam are displayed in Table 2 with the data broken down by institution and semester. These data illustrate the diversity of student performance profiles among our sample. Table 3 shows averages and counts broken down by the type of feedback (hereafter referred to as treatment) provided to the student between week 1 and week 2 testing. Our goal was to compare the different feedback treatments to address the research questions regarding how the types of feedback relate to student growth in exam performance and how that growth relates to students of different achievement levels and to different content areas.

Hierarchical linear modelling m2 coefficients and random effects

Using the results from pilot models (ESI section 3.3), the optimal model (m2) was used to model the full data to estimate the value added by each treatment. Since the primary focus of this study is the growth in student performance caused by each treatment, the initial ability levels (β_0 and $\theta_{0\ j(i)}$) are not reported here but are included in ESI section 5. When interpreting the coefficients directly, β_1 is the average growth of the full sample and $\theta_{1\ j(i)}$ is how much the growth of treatment j differs from that average. However, rather than interpreting β_1 and $\theta_{1j(i)}$ independently, a more useful metric is interpreting the sum of β_1 and $\theta_{1j(i)}$. The interpretation of the sum of these two coefficients is now the slope of treatment j, which can be interpreted as how many points of improvement were caused by treatment j. This model (m2) was run using dichotomous, open, and hierarchy scoring and the slopes ($\beta_1 + \theta_{1\ j(i)}$) for each treatment along with the standard error of $\theta_{1\ j(i)}$ are plotted in Figure 1.

The values of the slopes within each treatment vary between grading schemes. This is to be expected because the partial credit schemes gave students more opportunities for points which inflated their week 1 scores and gave them less room to improve during week 2. It is because of this week 1 inflation that dichotomous (lowest average score) is consistently the highest growth, followed by open (second lowest average score), followed by hierarchy (highest average score). These slopes may have agreed more closely with a more difficult exam where students wouldn't have been as likely to obtain the highest possible score.

Regardless of the slight differences in slopes within each treatment, the between-treatments trend is fairly consistent. The smallest gain was observed for the control treatment (delayed noncorrective only), whereas the largest gains were observed for the delayed elaborative feedback conditions (whether 1 or 2 answer attempts allowed) where students were asked to rework the problem or recall their initial response. All the scoring schemes showed the same progression of treatments with two exceptions under the hierarchy grading scheme (boxed in red on Figure 1). These differences were minor and only caused a flipping-in-order of two pairs of treatments.

Table 2 Week 1 descriptive statistics of exam performance for all grading schemes with the maximum possible score of 20.

Week 1		Institution					Overall Average
		I1	I2	I3	I4	I5	
Average Dichotomous Score	Fall	13.87	10.89	13.73	10.70		13.30
	Spring	12.59	11.05	13.63	10.97	14.56	13.09
Average Open Score	Fall	14.87	12.40	14.75	12.16		14.39
	Spring	13.74	12.55	14.72	12.39	15.52	14.22
Average Hierarchy Score	Fall	15.52	13.24	15.45	13.13		15.10
	Spring	14.52	13.44	15.38	13.35	16.18	14.96
							Overall Count
Count	Fall	379	85	417	74		955
	Spring	304	20	345	112	166	947
	Total	683	105	762	186	166	1902

Table 3 Comparison of week 1 and week 2 scores between treatments.

		Treatment											Overall Average
		1	2	3	4	5	6	7	8	9	10	11	
Average Dichotomous Score	**Week 1**	12.71	13.78	12.59	12.79	12.88	15.54	13.97	14.79	13.55	13.08	12.57	13.20
	Week 2	13.02	14.77	13.30	14.40	14.62	16.65	15.21	16.70	14.25	15.31	13.81	14.24
Average Open Score	**Week 1**	13.93	14.82	13.74	13.91	13.98	16.28	14.97	15.71	14.52	14.20	13.78	14.31
	Week 2	14.22	15.67	14.47	15.38	15.54	17.24	16.00	17.32	15.15	16.11	14.84	15.23
Average Hierarchy Score	**Week 1**	14.70	15.45	14.47	14.68	14.88	16.73	15.59	16.22	15.37	15.00	14.53	15.03
	Week 2	14.93	16.21	15.15	16.00	16.20	17.64	16.56	17.64	15.86	16.60	15.52	15.84
													Overall Count
Count		592	327	112	194	82	57	110	66	64	205	93	1902

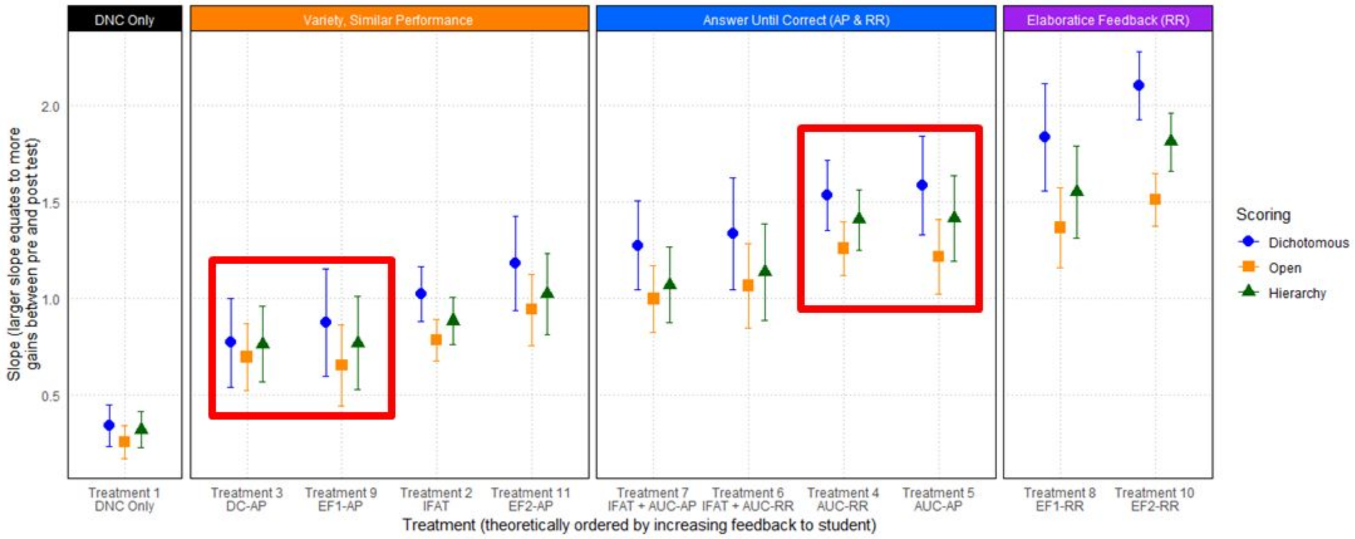


Fig. 1 Estimates for the slope ($\beta_1 + \theta_{1\ j(i)}$) of each treatment under the m2 model. The slope of each treatment is interpreted as how many points of improvement were caused by that treatment. Error bars correspond to the standard error of the treatment slope. Treatments are boxed based on how they were later collapsed.

Table 4 Order of treatments with the top leading to the most student improvement and the bottom leading to the least improvement. Colour coding of text corresponds to how the treatments were later collapsed.

Code	Treatment
10	**Delayed Elaborative Feedback (2 answer) (asked to rework/recall) and DNC (EF2-RR)**
8	**Delayed Elaborative Feedback (1 answer) (asked to rework/recall) and DNC (EF1-RR)**
5	**Delayed AUC no Elaborative Feedback (prior answer choices provided) and DNC (AUC-AP)**
4	**Delayed AUC no Elaborative Feedback (asked to rework/recall) and DNC (AUC-RR)**
6	**IFAT and Delayed AUC no Elaborative Feedback (asked to rework/recall) and DNC (IFAT + AUC-AP)**
7	**IFAT and Delayed AUC no Elaborative Feedback (prior answer choices provided) and DNC (IFAT + AUC-AP)**
11	**Delayed Elaborative Feedback (2 answer) (prior answer choices provided) and DNC (EF2-AP)**
2	**IFAT and DNC (IFAT)**
9	**Delayed Elaborative Feedback (1 answer) (prior answer choices provided) and DNC (EF1-AP)**
3	**Delayed Corrective no Elaborative Feedback (prior answer choices provided) and DNC (DC-AP)**
1	Delayed Noncorrective (DNC Only)

Sample collapse by treatment

While the order of student growth caused by the treatments is shown in Figure 1, many neighbouring treatments have standard error bars which overlap. In an attempt to further separate the differences caused by treatments and obtain sample sizes large enough to conduct IRT analysis, samples which received similar treatments were collapsed into groups. This process was guided by both the quantitative results that showed the ordering of treatment effectiveness, and by qualitatively analysing the treatments which are likely to produce similar results.

The first grouping (Group 1, purple in Table 4) was the highest performing treatment and consists of delayed elaborative feedback (asked to rework/recall). Grouping 2 (blue) was similar in that all treatments contained delayed AUC without elaborative feedback. The third grouping (orange) qualitatively has a wider variety of treatments together; however, quantitatively they all performed very similarly in Figure 1. Group 4 (black) was students who were only provided delayed noncorrective feedback and this treatment was left on its own because its standard error did not overlap with any of the other treatments, and because this treatment (as Treatment 1) served as the control for comparison.

After these treatment groupings (TG) were determined, a new HLM was constructed using a slight modification of the previous model (m2). The model for TG is $\mathbf{Y}_{ti} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 + \boldsymbol{\theta}_{0\ k(i)} + \boldsymbol{\theta}_{1\ k(i)} + \boldsymbol{\delta}_{0\ i} + \boldsymbol{\epsilon}_{ti}$ where all of the coefficients and indexes are interpreted in the same way. The only change to this model is that instead of treatment j, the new model has TG k. The values of the coefficients for this model are shown in ESI section 8. Figure 2 shows the slopes (β_1 + $\theta_{1\ k(i)}$) for each TG along with the group's standard errors. These TG for each scoring scheme (along with the true score which is discussed in the next section) show the expected ordering and separation of the treatment groups. Specifically, regardless of the scoring method, the smallest gain (slope) was observed for the control treatment (delayed noncorrective feedback), and the greatest gains were seen for delayed elaborative feedback where students were asked to recall or rework the problem.

Dichotomous 2-parameter logistic item response theory model

An additional benefit of collapsing the treatments into broader groupings is the increase in sample size of each group (see Table 5). Using these collapsed samples and within each week and TG, a 2-parameter logistic (2-PL) model was used which could accommodate both for question difficulty and discrimination. IRT true scores for each week were used following the same HLM procedure that was used above. These models, based on an alternative measure of student ability, led to the same conclusions that were drawn from the other dichotomous and polytomous scores (Figure 1). However, the true scores based on IRT modelling suggest the benefit caused by treatment group 2 and 3 may be understated by the other methods, and the benefit from treatment group 1 may be overstated (Figure 2).

Table 5 Sample size of each treatment and treatment grouping.

Code	Sample Size	Treatment Grouping Sample Size
10	205	Treatment Group 1 (TG1): Elaborative Feedback (RR) *N* = 271
8	66	
5	82	Treatment Group 2 (TG2): Answer Until Correct (AP & RR) *N* = 443
4	194	
6	57	
7	110	
11	93	Treatment Group 3 (TG3): Variety, Similar Performance *N* = 596
2	327	
9	64	
3	112	
1	592	Treatment Group 4 (TG4): DNC Only *N* = 592

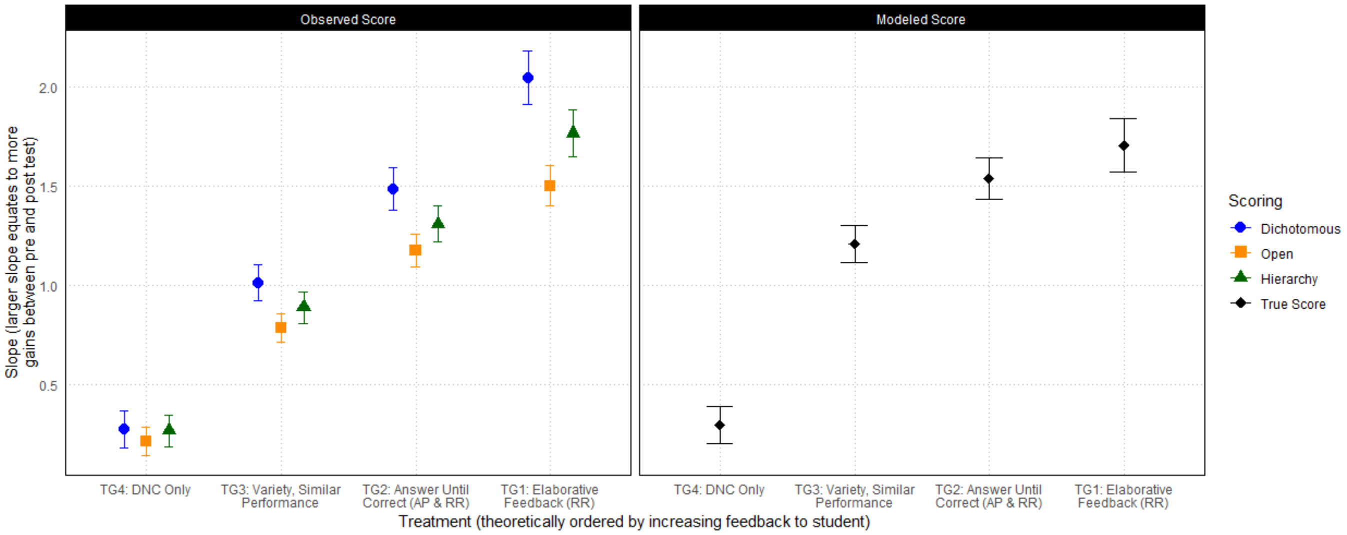


Fig. 2 Estimates for the slope (β_1 + $\theta_{1\ k(i)}$) of each treatment group under the m2 model. Error bars correspond to the standard error of the TG slope.

Table 6 Lord DIF detection under each Treatment Grouping (TG) with "X" indicating a significant difference between weeks at the 0.001 level.

Content Area	Question #	TG1	TG2	TG3	TG4
Chemical Formulas	Q16	X	X		
Conceptual Understanding of Molar Mass	Q14		X		
Empirical Formula	Q13	X	X	X	
Identify Excess Products	Q20	X	X		
Stoichiometric Calculations	Q8		X		
Limiting Reactant Calculations	Q19	X	X		
	Q9	X			
	Q10	X	X		

Lord's test for differential item functioning between week 1 and week 2 item response theory models

IRT analysis also allowed for week 1 and week 2 results to be more accurately compared through the use of Lord's Wald test. The significance threshold was set at the 0.001 level where a significant result indicates that week 1 and week 2 performed significantly different. Since the results of this test do not necessitate that week 2 performed better than week 1, ICC's (item characteristic curves) were analysed in the next section which show the probability of a student of any ability level answering the problem correctly. This analysis confirmed improvement in performance for most ability levels. Questions that showed a significant difference under each TG are shown in Table 6 and are later expanded upon in the section "Treatment grouping results summary". The presence of significant differences was then broken down by the content area of the question. Significance values for each Lord comparison can be found in ESI section 10.

Item Characteristic Curve (ICC) comparison

The primary limitation with the Lord results is while significant differences were found, no indication of effect size had been determined. While quantitative methods do exist for finding IRT overall effect size, for this study ICC's were compared to visualize how the effect size changed for each student ability level. In other words, ICC comparison allowed for visualization of not only which content areas benefited from each treatment, but also which student ability levels benefited the most. One example of direct ICC comparison is shown in Figure 3 with TG1, question 9. This figure shows that week 2 had a higher probability of students answering the question correctly for nearly every ability level. Students of average ability level (ability = 0), appear to benefit the most because this is where the week 1 and week 2 plots are most separated. It is also worth noting that the region where week 1 outperforms week 2 seems to be a result of the question discriminating better week 2 as opposed to a counterintuitive shift in improvement which benefits week 1. Comparing ICC's in this manner is effective; however, it is tedious and would lead to the analysis of 160 plots (20 questions x 2 weeks x 4 TG). Analysis can be dramatically simplified by instead investigating the difference between the weeks. This is shown for question 9 in the red box of Figure 3 and can be simply interpreted as the more positive the line, the more the treatment benefited students of that ability level. These differences for each question can then be summed to provide a rough estimate of which student abilities benefit overall under each TG. The plot of these sums is shown in Figure 4.

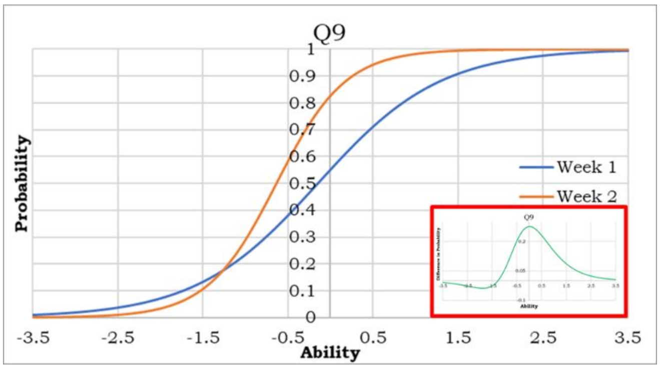


Fig. 3 Side-by-side comparison of ICCs between week 1 and week 2 of treatment grouping 1, question 9. Within the red box, is the area between these curves.

For ease of explanation, Figure 4 was broken into three categories of students using a cut-point of ±1.5: below-average (<–1.5 ability level), average (–1.5 to 1.5 ability level), and above-average (>1,5 ability level). The categories are colour coded in the figure. Interpretation of the plot shows that the TG effectiveness follows the expected pattern (TG1 > TG2 > TG3 > TG4) for average and above-average student. Above average students showed a smaller gap in growth due to the initial start of higher performance (ceiling effect). An interesting pattern is seen for below-average students (TG3 > TG2 > TG1 > TG4). This drop in treatment effectiveness may be caused by low-ability students not engaging in the 'asked to rework or recall' portions of the treatments which was pivotal for both TG1 and TG2. In addition, low-ability students may not engage as productively in "Answer Until Correct (AUC)" feedback methods, simply guessing through the answer choices rather than reasoning through alternate answers.

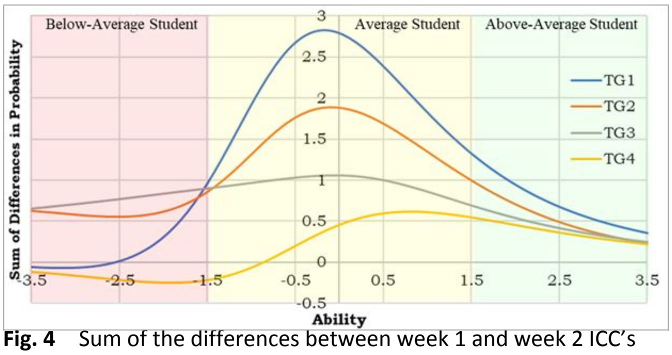


Fig. 4 Sum of the differences between week 1 and week 2 ICC's for every question within each treatment grouping.

The ICC can also be used to analyse content-specific growth. Raw differences in the ICC (as were used to establish Figure 4) after being broken down by content area can be found in ESI section 11. However, to better quantify the effect size of these differences, Cohen's h was calculated based on these ICC differences for each ability level. To determine if a content area experienced growth as a result of the TG, the cut-point of 0.5 was used which signifies a medium effect size (Cohen, 1988). Content groups that fell above this cut point are shown tabularly (Table 7) and analysed further in the section "Treatment grouping results summary".

Treatment grouping results summary

The benefit to students caused by each TG has now been analysed through two distinct methods: Lord's Wald test and ICC comparison. An independent qualitative approach to scoring was also completed (termed 'multimode scoring") and is documented in ESI section 12. A tabular summary of which content groups were determined to be benefited for each TG under each method are shown in Table 7.

Table 8 is a transformation of Table 7 and shows the ability level where students received the most benefit when ICCs within each content area were averaged. The ability levels are reported as a z-score so a value of 0 indicates average ability students, negative indicates below average ability students, and positive indicates above average students. For example, Table 8 shows that in the content area of "Chemical Formulas", TG1 resulted in the most growth for students with a factor score of -1.420. This means students who performed 1.420 standard deviations *below* the class average were most aided by this treatment grouping in this content area. However, this treatment grouping was most beneficial to students who were 1.036 standard deviations *above* the class average in the content areas of "Identify Excess Products". This transformation was only conducted for content areas where at least one of the three (Lord Statistic, IRT ICC, or Multimode) detection methods summarized in Table 7 showed growth for that treatment grouping. Interpretation of Table 8 is three-fold and involves analysis of the overall pattern, the within TG trend, and the between TG trend. Overall, the detection pattern is intuitive and shows that TG1 and TG2 led to growth in the most content areas, followed by TG3, and ending with no growth above the thresholds with TG4.

For analysis within treatment groupings, with the exception of a few content areas within TG1, the later the content area the higher the ability of students benefited. This means that lower achieving students are benefiting from the treatments by expanding their foundation knowledge of chemistry while higher achieving students who already possess sufficient foundational knowledge are expanding their peripheral knowledge.

Table 7 Full summary of improvement detection with X indicating improvement was observed.

Content	Measurement	TG1	TG2	TG3	TG4
Chemical Formulas	Lord Statistic	X	X		
	IRT ICC	X	X		
Conceptual Understanding of Molar Mass	Lord Statistic		X		
	IRT ICC	X		X	
Application of Molar Mass	No Significant Differences				
Mass Percent	Significant Differences Only Observed Through Multimode Analysis (See ESI section 12)				
Empirical Formula	Lord Statistic	X	X	X	
	IRT ICC	X	X		
Mole to Mole Ratio	No Significant Differences				
Identify Excess Products	Lord Statistic	X	X		
	IRT ICC	X			
Mole to Mole Conversion	No Significant Differences				
Stoichiometric Calculations	Lord Statistic		X		
	IRT ICC				
Limiting Reactant Calculations	Lord Statistic	X	X		
	IRT ICC	X			

When comparing between TG, the question becomes which grouping provides benefit to the widest array of students. Analysis of the ranges from Table 8 show that TG1 benefited students of the most diverse abilities with a range of -2.456 (-1.420 - 1.036 = -2.456). This was followed subsequently by TG2 (-1.689) and TG3 (-0.998). The large range of benefit to students provided by TG1 strengthens the argument for its use and confirms that it's not only beneficial to a specific ability student.

Table 8 Z-score ability that experienced the most growth for each treatment grouping where growth was seen. Negative values indicate peak growth for below average students while positive values indicate peak growth for above average students.

Content	TG1	TG2	TG3	TG4
Chemical Formulas	-1.420	-1.523		
Conceptual Understanding of Molar Mass	-0.787	-1.247	-0.883	
Mass Percent	-0.269			
Empirical Formula	-0.422	-0.806	0.115	
Identify Excess Products	1.036	-0.576		
Stoichiometric Calculations		-0.192		
Limiting Reactant Calculations	-0.294	0.166		

Conclusions

Implications

The results provided by the m2 HLM show a clear ordering of the effectiveness of different types of feedback treatments (research question 1). After sample collapse by treatment, the standard errors did not overlap with neighbouring groupings and showed particular benefit for providing students with delayed elaborative feedback coupled with students being asked to recall or rework each problem. Further investigation into this growth showed that students of a wide range of abilities all benefited from TG1 (delayed elaborative feedback) lending weight to its use for a larger variety of students (research question 2). More specifically, the data presented here suggest practitioners will provide the greatest benefit to their students, as a whole, by providing them with an opportunity for delayed elaborative feedback along with having the students recall/rework the exam questions. This can be operationalized by providing students with the opportunity to rework exam questions with item-by-item feedback, for example with an electronic platform or course management system which would provide elaborative feedback in real time. Note this suggestion does not necessarily advocate for exam retakes for improved exam grades, but as a post-exam assignment with the purpose of learning from errors for future assessments.

It is noteworthy to address the significant instructor time investment of programming in item-by-item elaborative feedback. In the case where such feedback is not realistic to be implemented, dramatic student improvement was also seen using AUC (TG2). Assuming the course-management system supports AUC, this method requires no additional instructor intervention and therefore may be a more preferable option to many instructors.

Another important result of the work presented here is that with all methods of analysis, the control treatment, which consisted of delayed noncorrective feedback that provided a score but no information on missed questions or correct answers, showed the lowest gain, and as shown in Tables 7 and 8 (as TG4) showed no gain for any content area. Further, delayed corrective feedback (Treatment 3 in Table 1, collapsed into TG3 in the subsequent analysis) showed little to no gain in performance either. This feedback condition mimics the predominant feedback given in a large-enrolment course: students can see which questions they answered correctly/incorrectly and an answer key is posted. The work presented here indicates that this method of feedback does not lead to improved performance for students. Effective feedback must include an opportunity to rework exam questions through initial attempts and/or through answer-until-correct attempts (TG1 & TG2), and the most effective feedback will also contain item-by-item (TG1) elaborative feedback. The results of this work have implications for best practices in exam feedback as well as for online homework systems.

An interesting caveat is that lower-performing students experience the lowest gain from the delayed elaborative feedback. Thus, if an instructor is attempting to help primarily low-performing students, it may be more beneficial to provide those students with their original answers to questions before they engage in a delayed feedback opportunity as opposed to being asked to recall or rework questions as is indicated in Figure 4. It is also important to note that for reasonable elaborative feedback to be given and for student improvement through item-by-item feedback, multiple-choice assessments should be built with logical distractors. There is also an advantage to incorporating partial credit scoring, which provides credit for partially-correct processes, a method of giving feedback to students with initial scoring (Murphy et al.).

These results also indicated a relationship between feedback type and content area being assessed (research question 3). However, while this analysis shows that a relationship between content and feedback exists, this study was not designed to make claims about what causes a particular type of feedback to be effective within a specific content area. Based on prior research, it can be assumed a portion of this relationship can be explained by the ratio of complex problem solving versus fact-based retrieval the item required (Wheeler et al., 2003; Roediger and Karpicke, 2006; Coppens et al., 2011; Karpicke and Aue, 2015; Van Gog and Sweller, 2015; Van Gog et al., 2015). However, a deeper investigation is required to better understand what types of problems are ideal for which type of feedback.

Regardless of the type of feedback an instructor *wishes* to provide to students, an important consideration is the feedback options provided by their course management system. For example, functionality to provide students with immediate feedback during testing or testing feedback (as opposed to feedback following test completion) may not be available. The feedback mechanisms we investigated follow best practices of connecting item stem with answer choices to student answers to corrective feedback with feedback coming at the moment students are thinking about a particular item. Most online course managements systems only allow for whole test feedback at the end of submission. Given that the benefit of different feedback types was found to vary based on content and student ability, we encourage course management systems to provide a greater variety of possible feedback mechanisms for instructors to choose from.

Limitations

The effectiveness of each individual treatment is based on student performance in introductory chemistry and may not be consistent at other levels of chemistry or in other fields. Additionally, the influence of TG on content-specific growth uses a limited number of questions to predict student growth for a more general content area. The intention of the content-specific growth results is not to encourage a specific type of feedback based on the content area being studied. Rather, this study only demonstrates that students may respond to feedback differently based on the content being assessed. To understand the cause of differential improvement between content areas, a deeper qualitative investigation into changes in student process would be necessary.

The differential performance of treatments between content areas would be expected to be even more apparent for content which rely on different processes. Another limitation of this work is that many of the items tested could be solved algorithmically. The Testing Effect literature suggests the Testing Effect may be even stronger for fact-based items than the complex problem-solving items we tested (Karpicke and Aue, 2015; Van Gog and Sweller, 2015; Van Gog et al., 2015).

Finally, while these results are based on 5 distinct and diverse universities, they do not encompass or account for all populations of students, and no analysis was done to date on this data to investigate how different student populations (besides ability) grow differently when exposed to these treatments. Furthermore, even though student growth was observed between the week 1 and week 2 assessment, it is unclear how well the learning gains students experienced translated to other course work later in the semester. Similar feedback research in the more authentic classroom setting is an important next step.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Funding from NSF TUES grant (DUE 1140914), lead-PI Jamie Schneider, and NSF IUSE Grant (DUE 1625233), PI Jamie Schneider. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We are also grateful to Dr. Christopher Bauer and Dr. Panayiota Kendeou for their continued guidance and assessment of our projects as the external evaluators on the NSF grants. Dr. Arunendu Chatterjee provided much statistical advice and preliminary analysis during the early stages of the project. We would also like to acknowledge many undergraduate research students (Andrew Beszhak, Alex Cimaglio, Marcus Engle, Chad Fleck, Julian Ikeri, Matthew Larson) who helped with data entry and organization. Finally, we would like to show our appreciation to all the students who took the exams and the professors who aided in the distribution of the exams.

Notes and references

‡ Now located at: University of Wisconsin-Madison, Madison, Wisconsin 53706

§ Now located at: Auburn University, Auburn, Alabama 36849

American Psychological Association (APA), (2021), *APA Dictionary of Psychology*.

Andaya G., Hrabak V. D., Reyes S. T., Diaz R. E., and McDonald K. K., (2017), Examining the Effectiveness of a Postexam Review Activity to Promote Self-Regulation in Introductory Biology Students. *J Coll Sci Teach*, **46**(4), 84–92.

Bates D., Mächler M., Bolker B., and Walker S., (2015), Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw*, **67**(1), 1–48, DOI: 10.18637/jss.v067.i01.

Butler A. C., Godbole N., and Marsh E. J., (2013), Explanation feedback is better than correct answer feedback for promoting transfer of learning. *J Educ Psychol*, **105**(2), 290–298, DOI: 10.1037/a0031026.

Butler A. C., Karpicke J. D., and Roediger H. L., (2007), The Effect of Type and Timing of Feedback on Learning From Multiple-Choice Tests. *J Exp Psychol Appl*, **13**(4), 273–281, DOI: 10.1037/1076-898X.13.4.273.

Butler A. C. and Roediger H. L., (2008), Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Mem Cognit*, **36**(3), 604–616, DOI: 10.3758/MC.36.3.604.

Butler A. C. and Woodward N. R., (2018), Toward consilience in the use of task-level feedback to promote learning. *Psychology of Learning and Motivation - Advances in Research and Theory*, **69**, 1–38, DOI: 10.1016/BS.PLM.2018.09.001.

Cohen J., (1988), Statistical Power Analysis for the Behavioural Science (2nd Edition).

Coppens L. C., Verkoeijen P. P. J. L., and Rikers R. M. J. P., (2011), Learning Adinkra symbols: The effect of testing. *Journal of Cognitive Psychology*, **23**(3), 351–357, DOI: 10.1080/20445911.2011.507188.

Dragulescu A. and Arendt C., (2020), xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. *CRAN*.

Epstein Educational Enterprises, Immediate Feedback Assessment Technique (IF-AT). Center for the Enhancement of Teaching & Learning, (Figure 2), 1–3.

Gelman A. and Su Y.-S., (2020), arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. *CRAN*.

Gravic, (2023), Remark.

Henderson C. and Harper K. A., (2009), Quiz Corrections: Improving Learning by Encouraging Students to Reflect on Their Mistakes. *Phys Teach*, **47**(9), 581–586, DOI: 10.1119/1.3264589.

Hintzman D. L., (2010), How does repetition affect memory? Evidence from judgments of recency. *Mem Cognit*, **38**(1), 102–115, DOI: 10.3758/MC.38.1.102.

Karpicke J. D., (2012), Retrieval-Based Learning: Active Retrieval Promotes Meaningful Learning. *Curr Dir Psychol Sci*, **21**(3), 157–163, DOI: 10.1177/0963721412443552.

Karpicke J. D. and Aue W. R., (2015), The Testing Effect Is Alive and Well with Complex Materials. *Educ Psychol Rev*, **27**(2), 317–326, DOI: 10.1007/s10648-015-9309-3.

Karpicke J. D. and Roediger H. L., (2008), The critical importance of retrieval for learning. *Science*, **319**(5865), 966–968, DOI: 10.1126/science.1152408.

Kulhavy R. W. and Anderson R. C., (1972), Delay-retention effect with multiple-choice tests. *J Educ Psychol*, **63**(5), 505–512, DOI: 10.1037/h0033243.

Magis D., Béland S., Tuerlinckx F., and De Boeck P., (2010), A general framework and an R package for the detection of dichotomous differential item functioning. *Behav Res Methods*, **42**(3), 847–862, DOI: 10.3758/BRM.42.3.847.

Moreno R., (2004), Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instr Sci*, **32**(1–2), 99–113, DOI: 10.1023/b:truc.0000021811.66966.1d.

Mullet H. G., Butler A. C., Verdin B., von Borries R., and Marsh E. J., (2014), Delaying feedback promotes transfer of knowledge despite student preferences to receive feedback immediately. *J*

Appl Res Mem Cogn, **3**(3), 222–229, DOI: 10.1016/j.jarmac.2014.05.001.

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.

Knaus K. J., Murphy K. L., and Holme T. A., (2009), Designing Chemistry Practice Exams for Enhanced Benefits. An Instrument for Comparing Performance and Mental Effort Measures. *J Chem Educ*, **86**(7), 827–832, DOI: 10.1021/ed086p827.

Pinheiro J., Douglas B., DebRoy S., Sarkar D., and R Core Team, (2020), nlme.

R Core Team, (2022), R: A Language and Environment for Statistical Computing.

Risley J. M., (2007), Reworking Exams To Teach Chemistry Content and Reinforce Student Learning. *J Chem Educ*, **84**(9), 1445, DOI: 10.1021/ed084p1445.

Rizopoulos D., (2006), ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *J Stat Softw*, **17**(5), 1–25, DOI: 10.18637/jss.v017.i05.

Roediger H. L. and Karpicke J. D., (2006), Test-enhanced learning: Taking memory tests improves long-term retention. *Psychol Sci*, **17**(3), 249–255, DOI: 10.1111/j.1467-9280.2006.01693.x.

Roediger H. L. and Marsh E. J., (2005), The Positive and Negative Consequences of Multiple-Choice Testing. *J Exp Psychol Learn Mem Cogn*, **31**(5), 1155–1159, DOI: 10.1037/0278-7393.31.5.1155.

Rowland C. A., (2014), The Effect of Testing Versus Restudy on Retention: A Meta-Analytic Review of the Testing Effect. *Psychol Bull*, **140**(6), 1432–1463, DOI: 10.1037/a0037559.

Schneider J. L., Hein S. M., and Murphy K. L., (2014), Feedback in testing, the missing link. ACS Symposium Series, 1182, 93–112, DOI: 10.1021/bk-2014-1182.ch006.

Schreurs D., Trate J., Srinivasan S., Teichert M., Luxford C., Schneider J., and Murphy K., (2024), Investigation into the intersection between response process validity and answer-until-correct validity: Development of the Repeated Attempt Processing Issue Detection (RAPID) method. *Chemistry Education Research and Practice*.

Schurmeier K. D., Atwood C. H., Shepler C. G., and Lautenschlager G. J., (2010), Using item response theory to assess changes in student performance based on changes in question wording. *J Chem Educ*, **87**(11), 1268–1272, DOI: 10.1021/ed100422c.

Todd K., Therriault D. J., and Angerhofer A., (2021), Improving students' summative knowledge of introductory chemistry through the forward testing effect: examining the role of retrieval practice quizzing. *Chemistry Education Research and Practice*, **22**(1), 175–181, DOI: 10.1039/d0rp00185f.

Trate J. M., Teichert M. A., Murphy K. L., Srinivasan S., Luxford C. J., and Schneider J. L., (2020), Remote Interview Methods in Chemical Education Research. *J Chem Educ*, **97**(9), 2421–2429, DOI: 10.1021/acs.jchemed.0c00680.

Van Gog T., Kester L., Dirkx K., Hoogerheide V., Boerboom J., and Verkoeijen P. P. J. L., (2015), Testing After Worked Example Study Does Not Enhance Delayed Problem-Solving Performance Compared to Restudy. *Educ Psychol Rev*, **27**(2), 265–289, DOI: 10.1007/s10648-015-9297-3.

Van Gog T. and Sweller J., (2015), Not New, but Nearly Forgotten: the Testing Effect Decreases or even Disappears as the Complexity of Learning Materials Increases. *Educ Psychol Rev*, **27**(2), 247–264, DOI: 10.1007/s10648-015-9310-x.

Wheeler M. A., Ewers M., and Buonanno J. F., (2003), Different rates of forgetting following study versus test trials. *Memory*, **11**(6), 571–580, DOI: 10.1080/09658210244000414.

Wickham H. and Bryan J., (2019), readxl: Read Excel Files.

Woltman H., Feldstain A., Mackay J. C., Rocchi M., Woltman H., Feldstain A., and Rocchi M., (2012), An introduction to hierarchical linear modeling. **8**(1), 52–69.

Wright J. H. and Gescheider G. A., (1970), Role of Immediate and Delayed Knowledge of Results in Paired-Associate Learning under the Anticipation Procedure. *J Psychol*, **74**(2), 249–257, DOI: 10.1080/00223980.1970.9923736.

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)

ESI 2

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

ESI 3

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

ESI 4

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

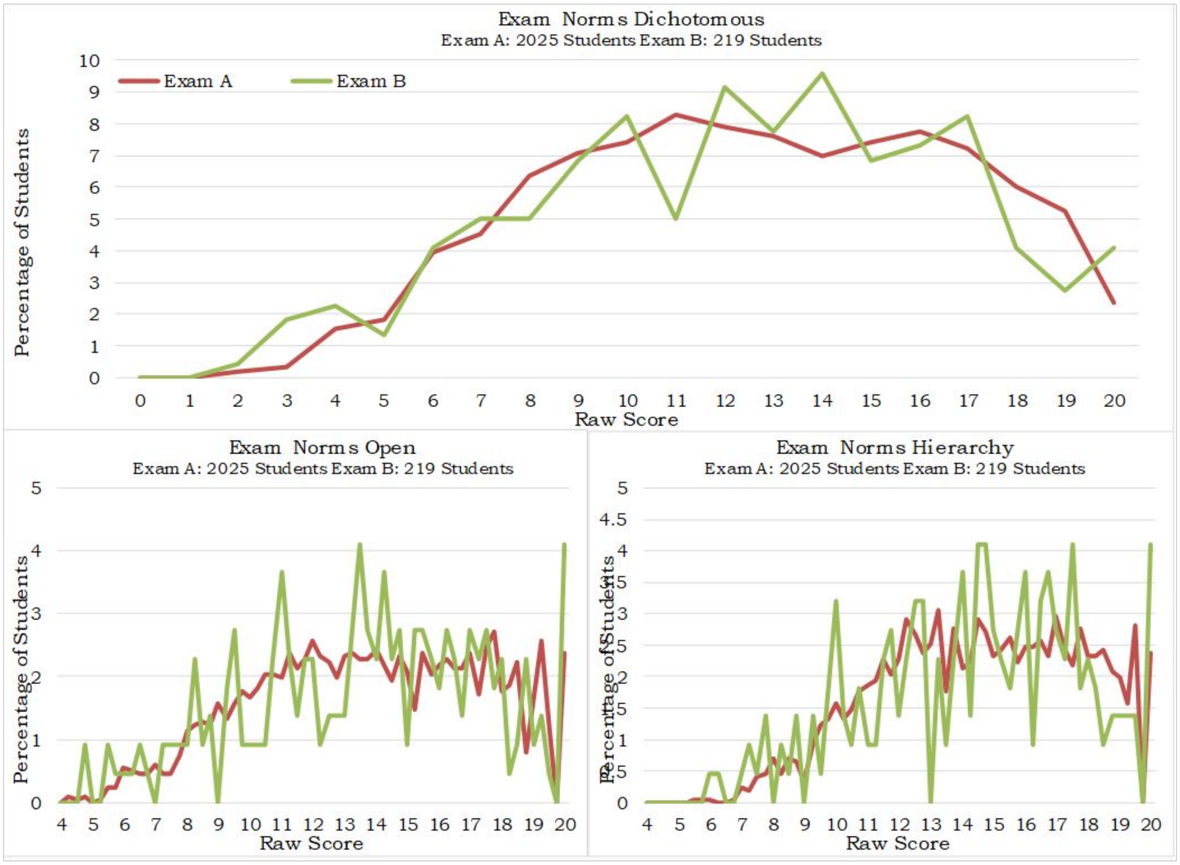


Fig. 1 Percent of each raw score obtained on Exam A and Exam B under each of the grading methods.

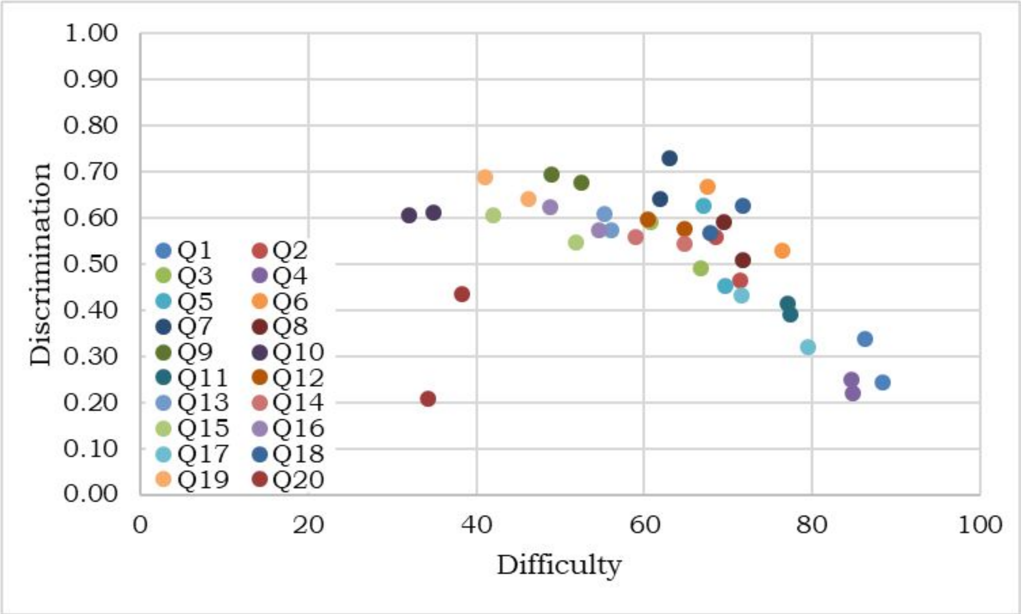


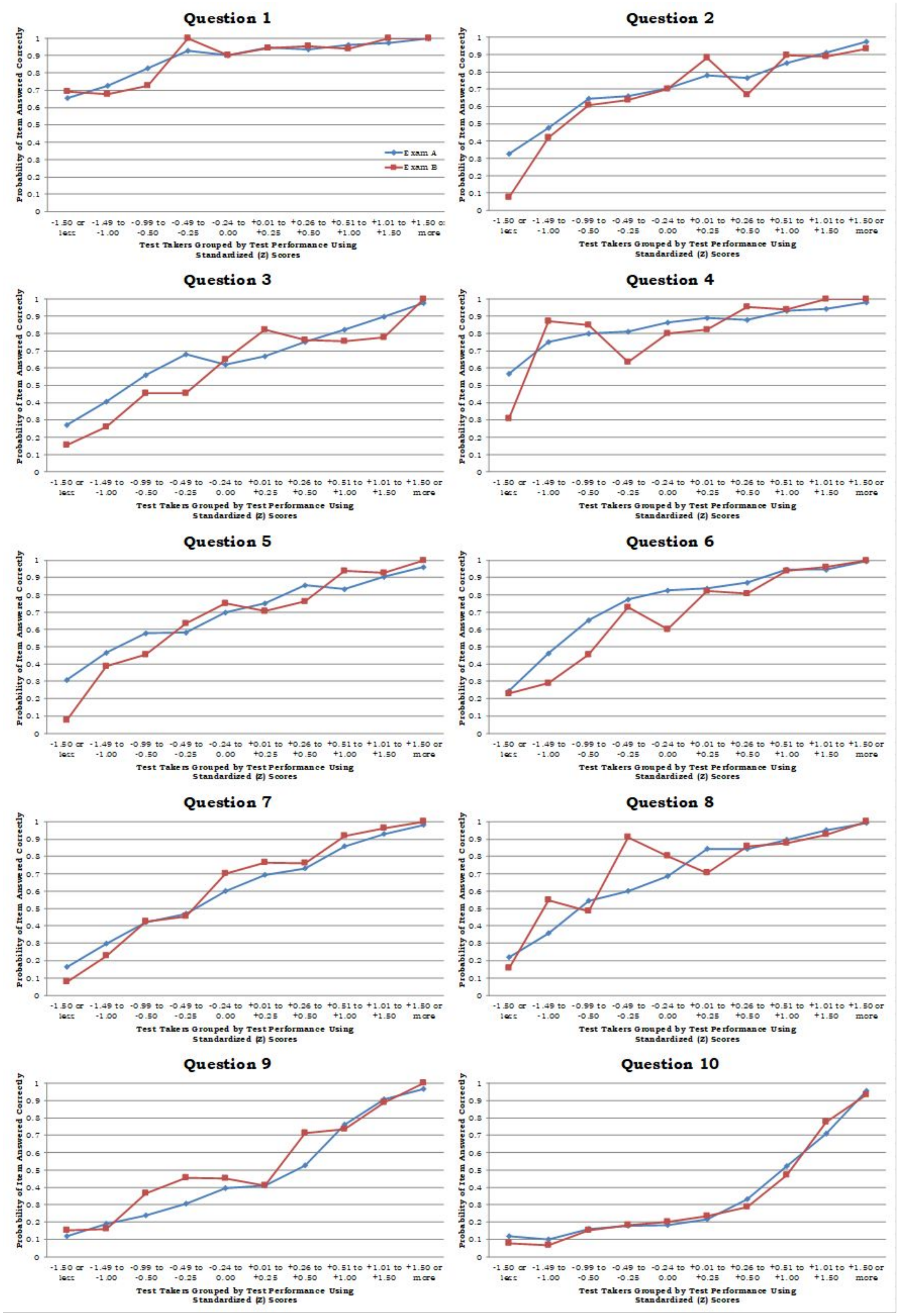
Fig. 2 Plot of item difficulty versus discrimination. Each question has two points, one from exam A and one from exam B.

ESI 6

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

ESI 7


Question 1
Question 2
Question 3
Question 4
Question 5
Question 6
Question 7
Question 8
Question 9
Question 10
Probability of Item Answered Correctly
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-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
Exam A
Exam B

ESI 8

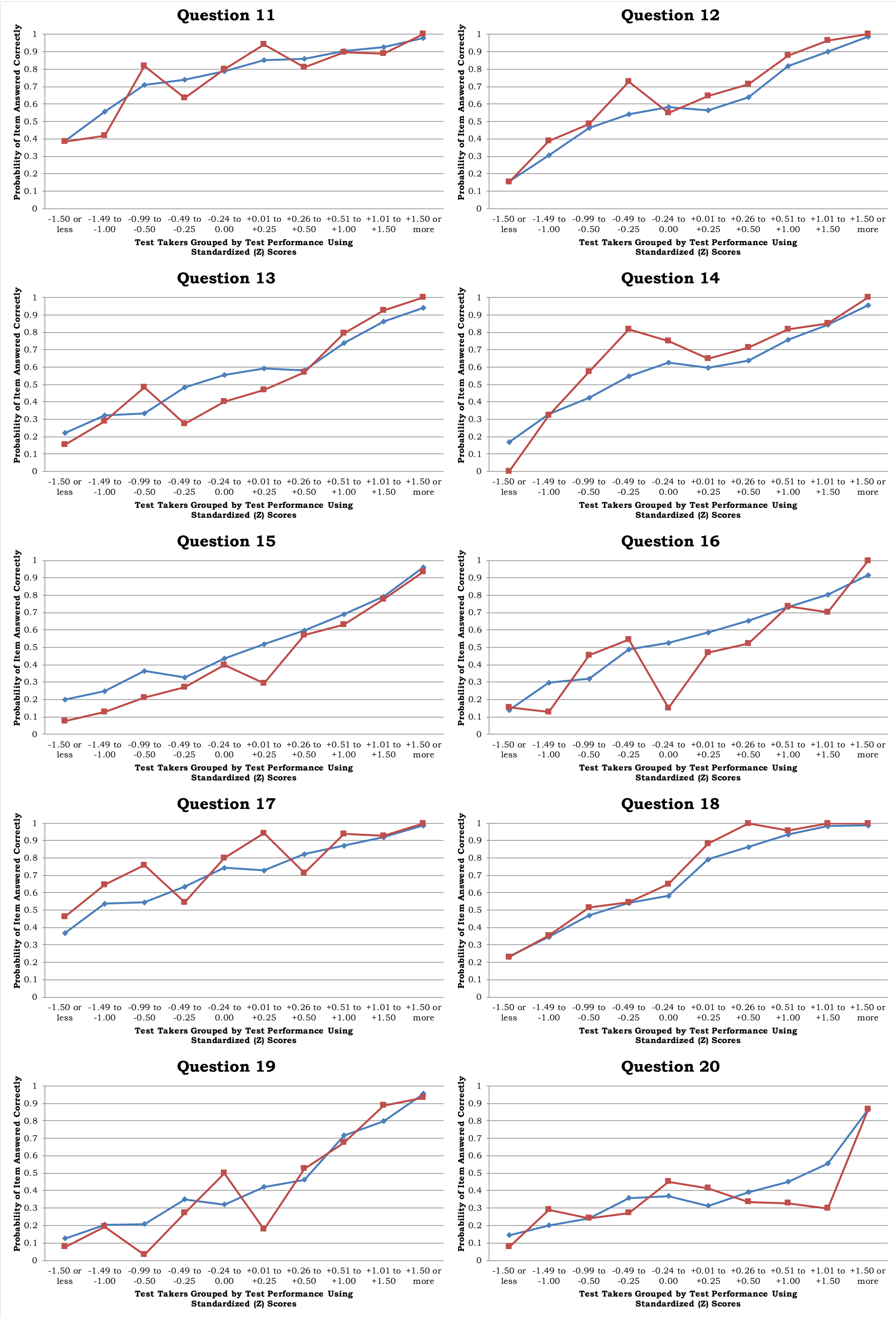


Fig. 3 Item plots of each question for Exam A (blue) and Exam B (red) showing similar performance. Item plots were constructed in the same manner as discussed by Holme (Holme and Murphy, 2011).

ESI 9

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

ESI 10

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.

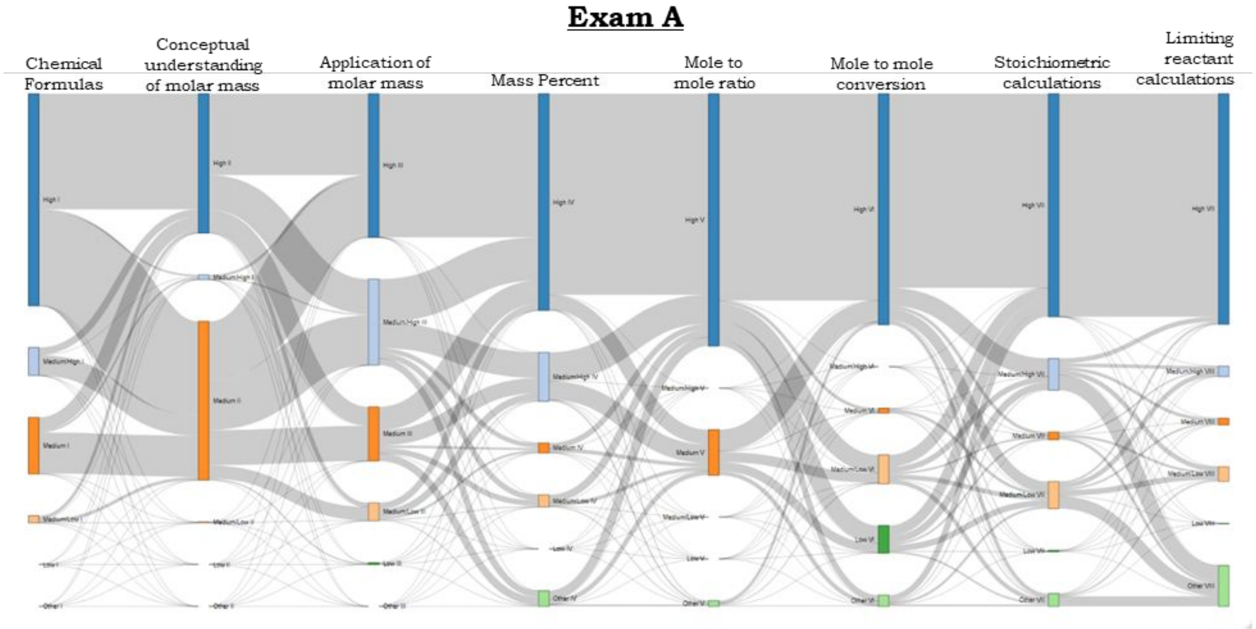


Fig. 4 Sankey diagram for exam A showing student categorization (high, medium/high, medium, medium/low, low, or other) and movement between predicted ability levels within content areas.

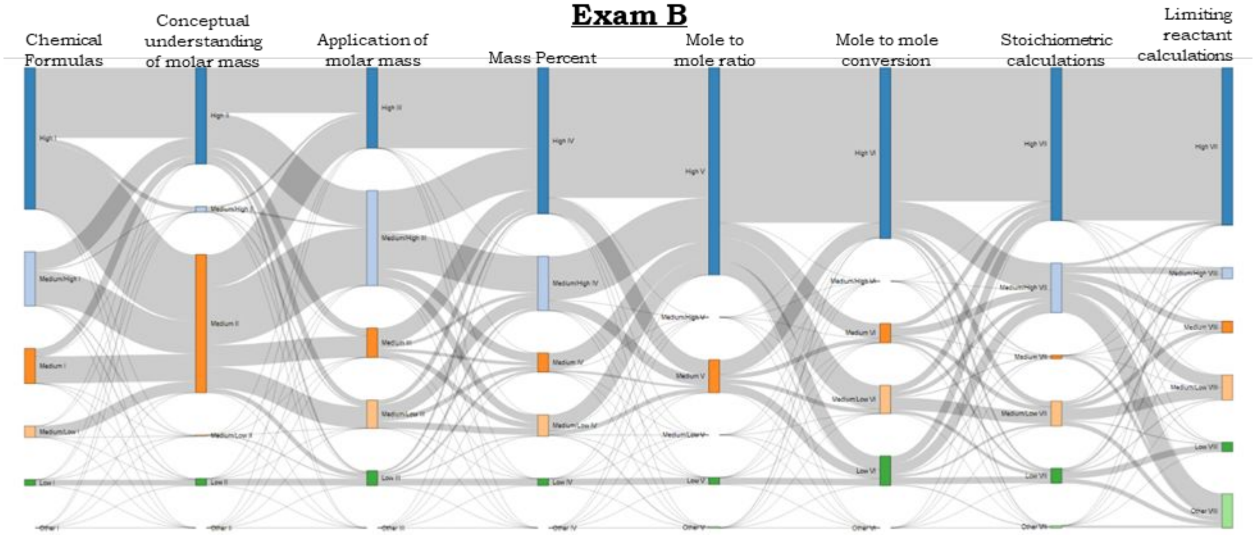


Fig. 5 Sankey diagram for exam B showing student categorization (high, medium/high, medium, medium/low, low, or other) and movement between predicted ability levels within content areas.

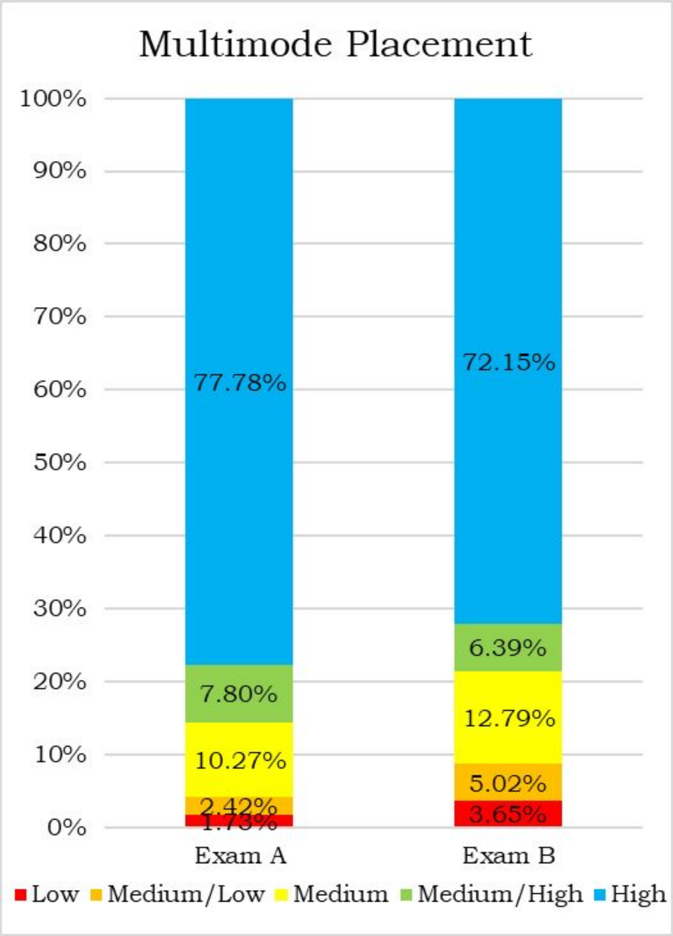


Fig. 6 Predicted overall ability of students based on multimode results.

ESI 12

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

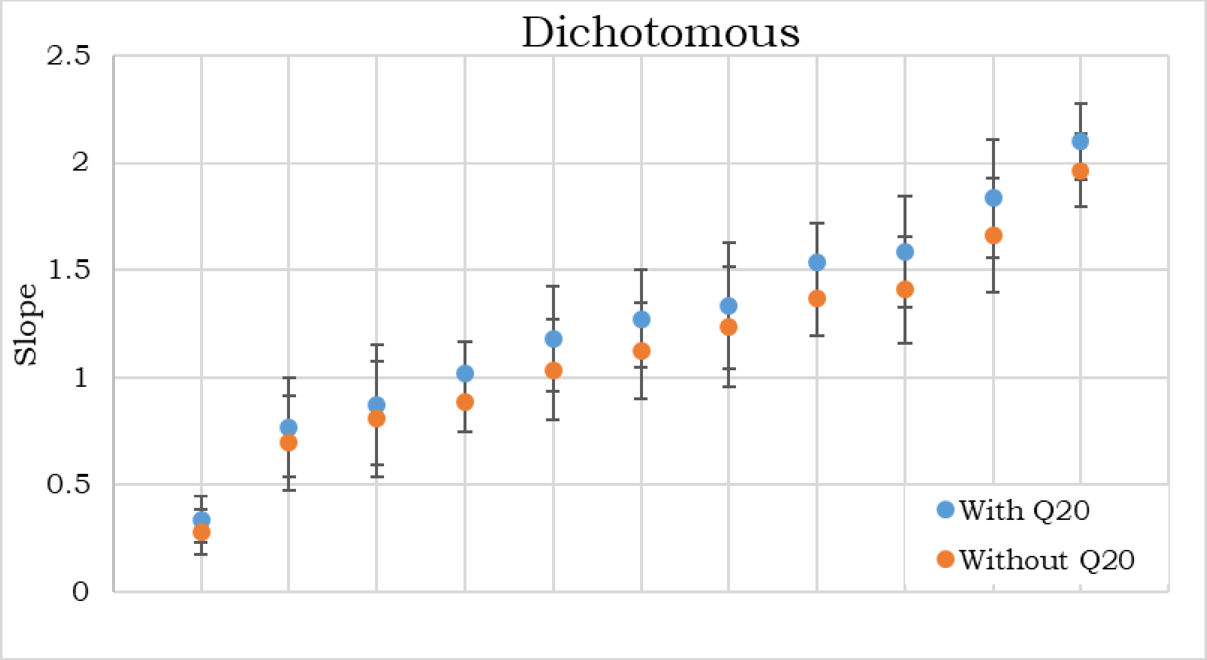


Fig. 7 Dichotomous student growth, along with the standard error, caused by each treatment both with and without Q20.

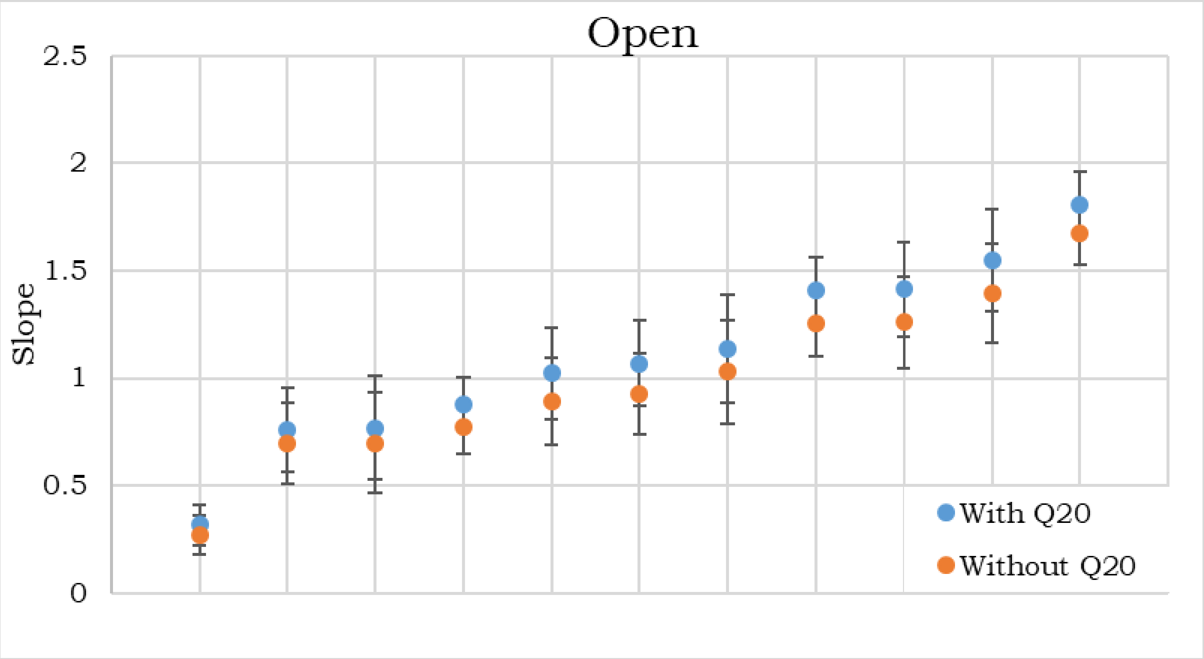


Fig 8 Open student growth, along with the standard error, caused by each treatment both with and without Q20.

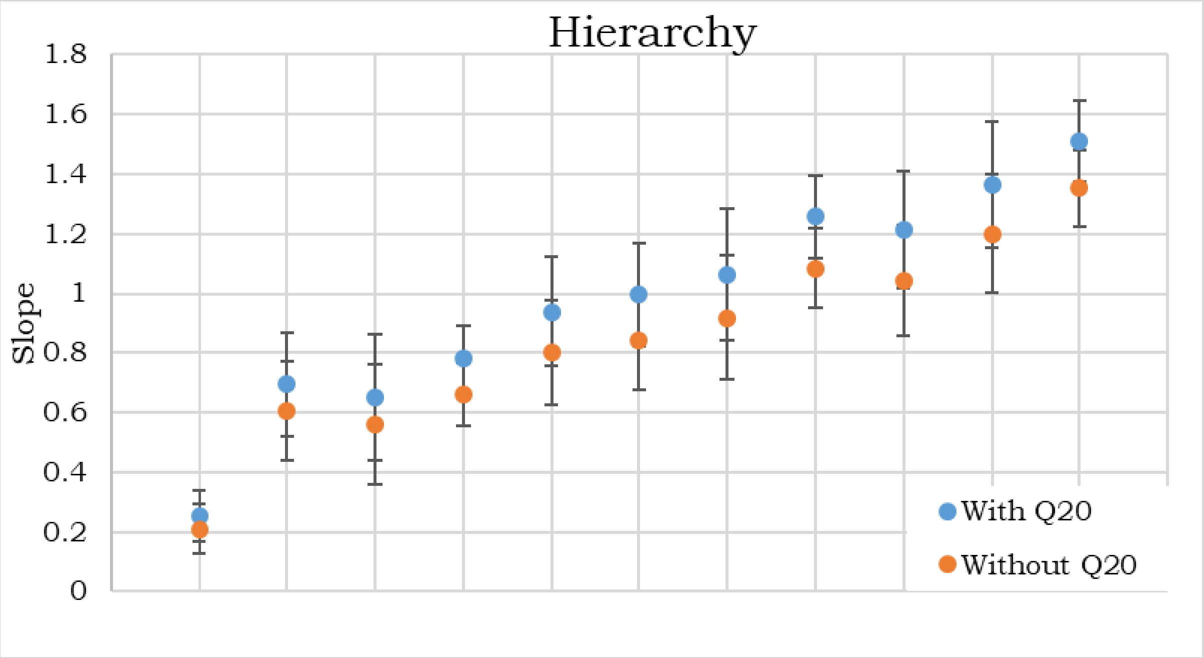
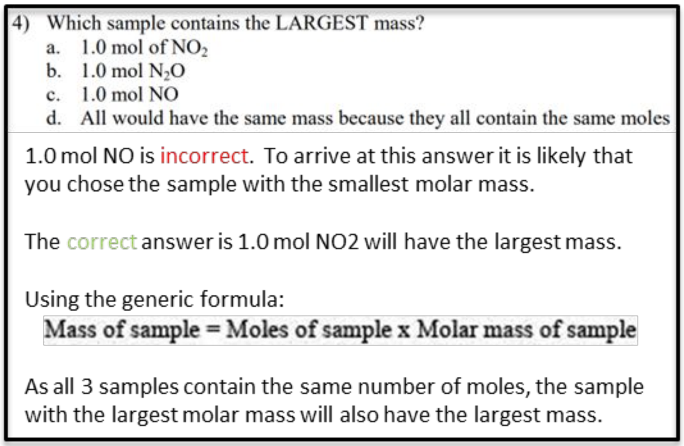


Fig. 8 Hierarchy student growth, along with the standard error, caused by each treatment both with and without Q20.

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

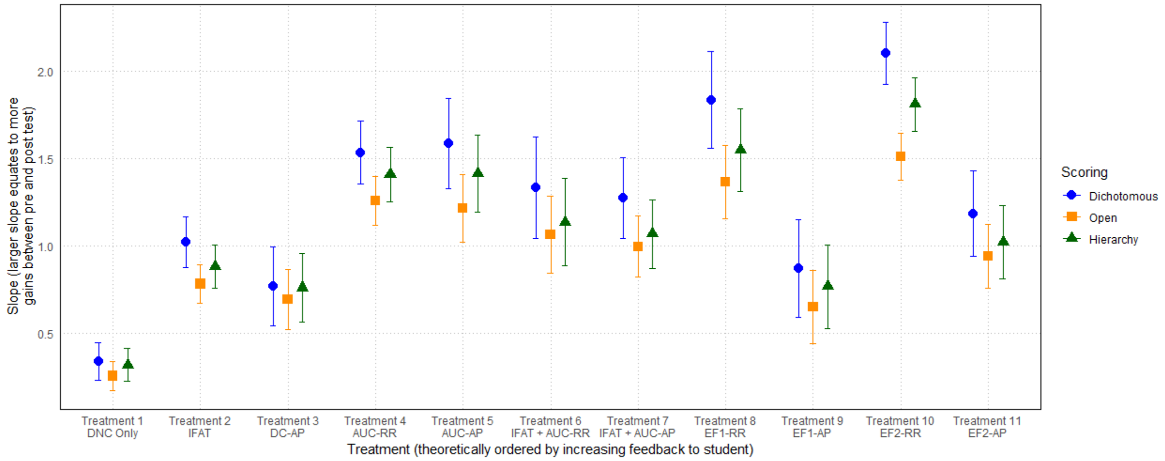


Fig. 10 Estimates for the slope ($\beta_1 + \theta_{1\ j(i)}$) of each treatment under the m2 model. The slope of each treatment is interpreted as how many points of improvement were caused by that treatment. Error bars correspond to the standard error of the treatment slope. Plot is functionally identical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

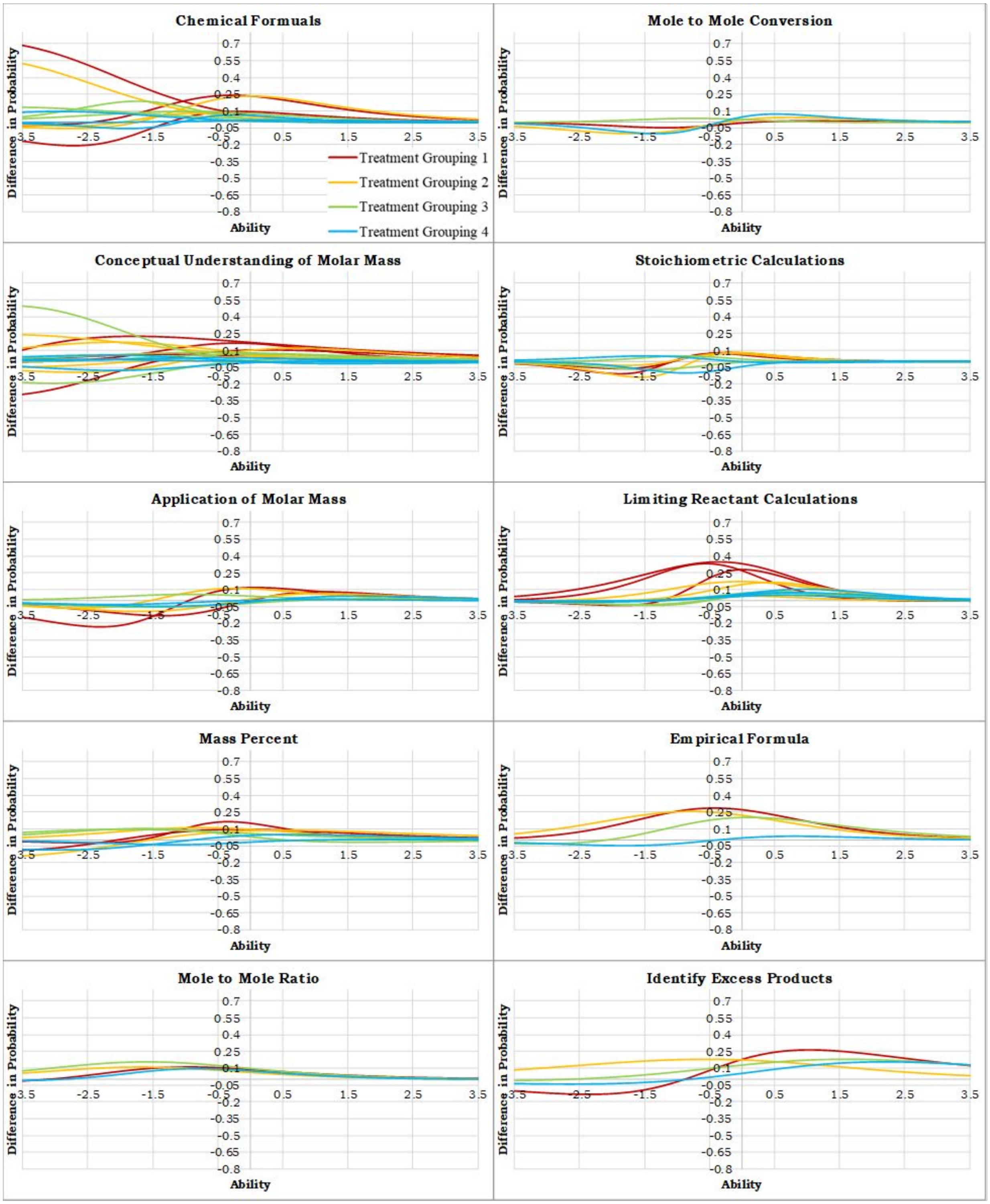


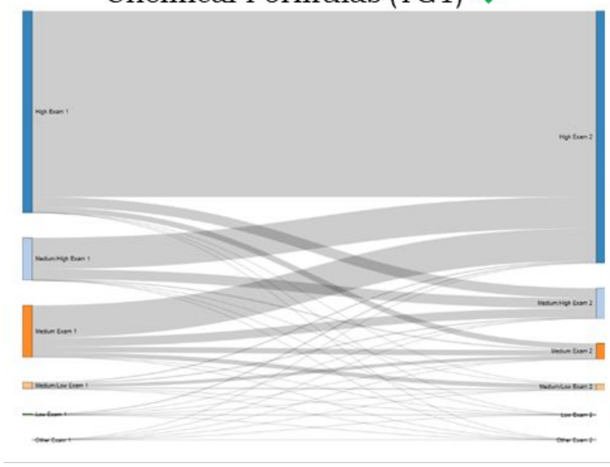
Fig. 11 Difference between week 1 and week 2 ICC's plotted within content areas.

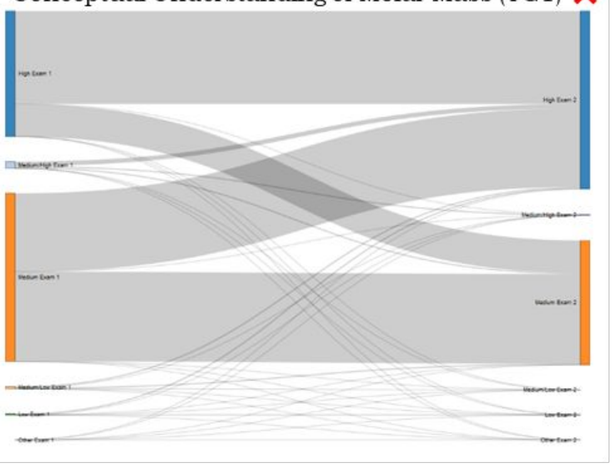
12) Multimode analysis

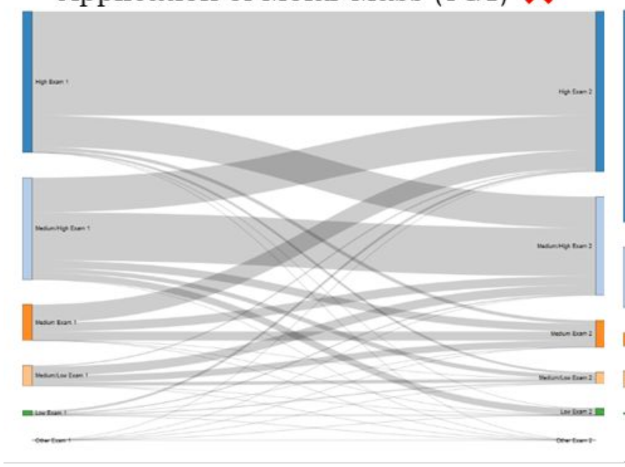
In addition to the quantitative methods used to assess each treatment, qualitative measures were also used to access student growth. Multimode grading was previously conducted on this exam, and the results of this grading scheme were also applied to the research questions here-in (Murphy et al.). As 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.

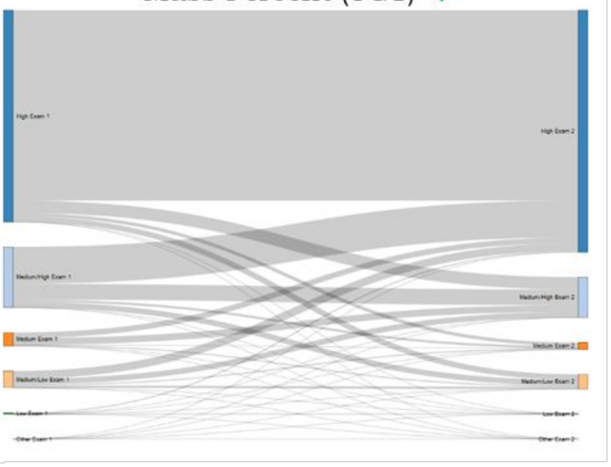
12.1) Content-specific multimode analysis

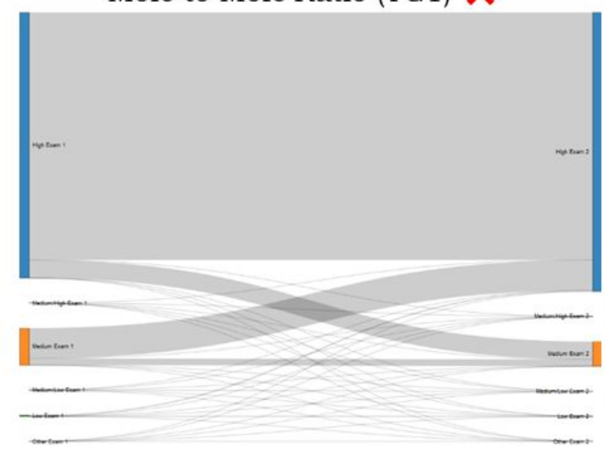
Students' ability within each content area each week was determined using the multimode method (Murphy et al.). After content-specific ability levels were determined, Sankey diagrams were constructed to visualize ability migration from week 1 (left column) to week 2 (right column). The height of each 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.

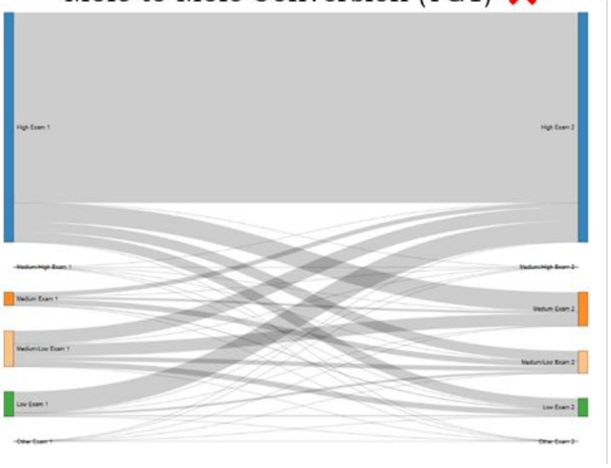

Chemical Formulas (TG1) ✓


Conceptual Understanding of Molar Mass (TG1) ✗

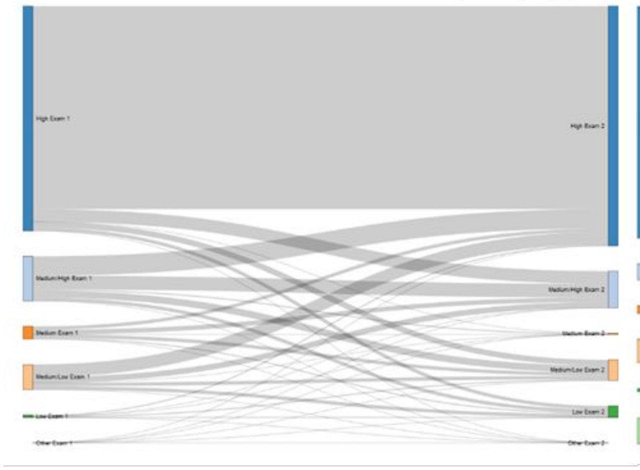

Application of Molar Mass (TG1) ✗


Mass Percent (TG1) ✓

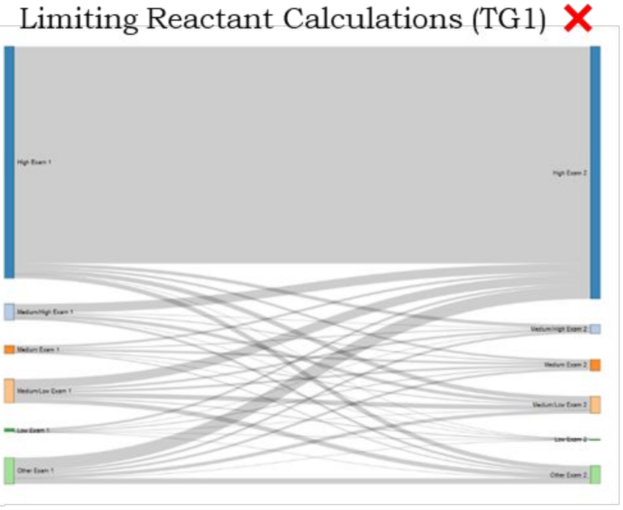

Mole to Mole Ratio (TG1) ✗


Mole to Mole Conversion (TG1) ✗

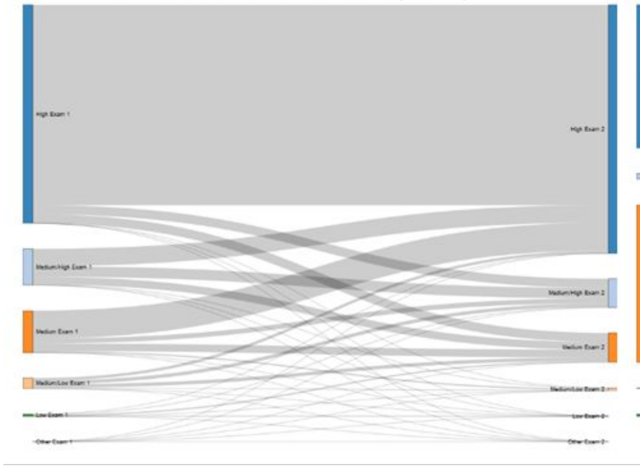
Stoichiometric Calculations (TG1) ✗



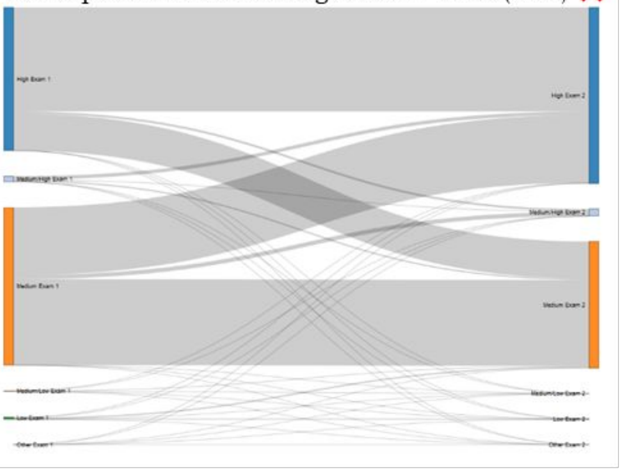
Limiting Reactant Calculations (TG1) ✗



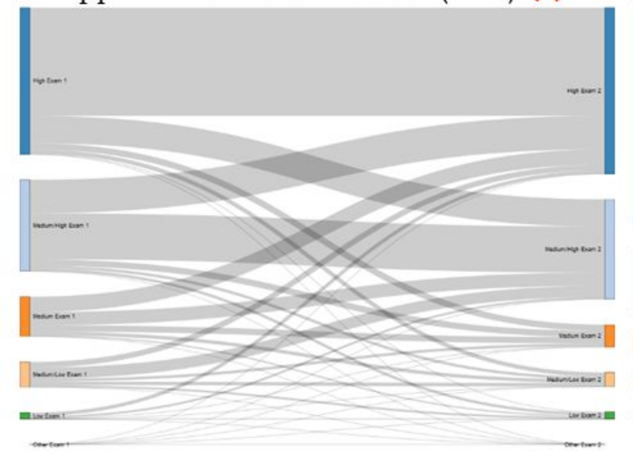
Chemical Formulas (TG2) ✓



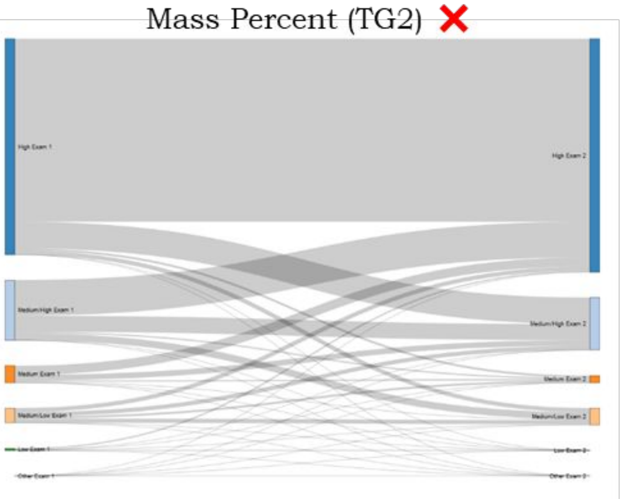
Conceptual Understanding of Molar Mass (TG2) ✗



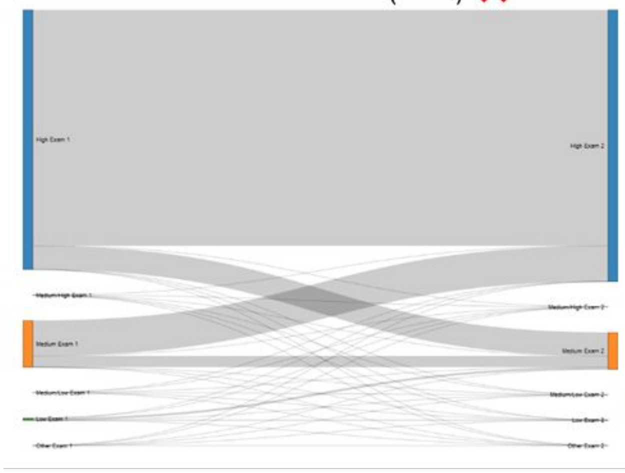
Application of Molar Mass (TG2) ✗



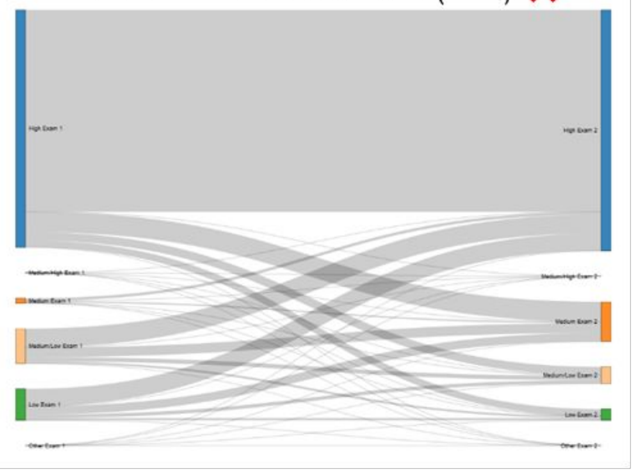
Mass Percent (TG2) ✗



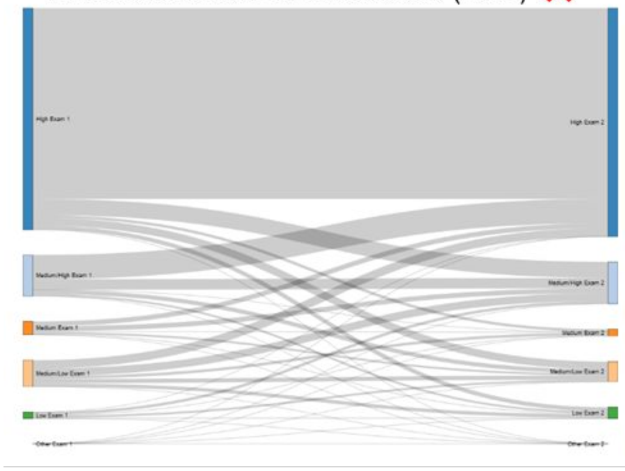
Mole to Mole Ratio (TG2) ✗



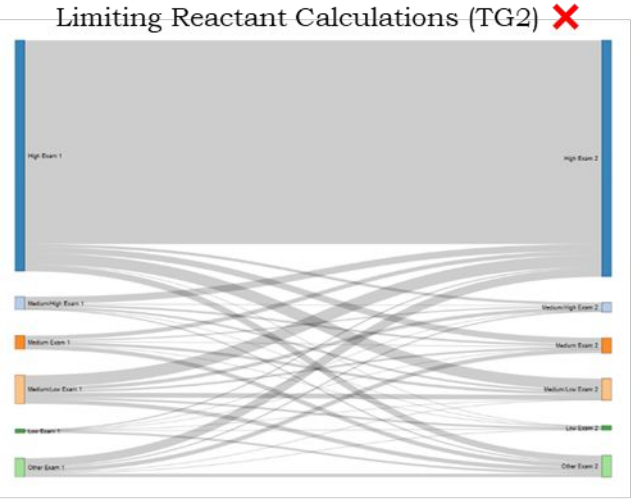
Mole to Mole Conversion (TG2) ✗



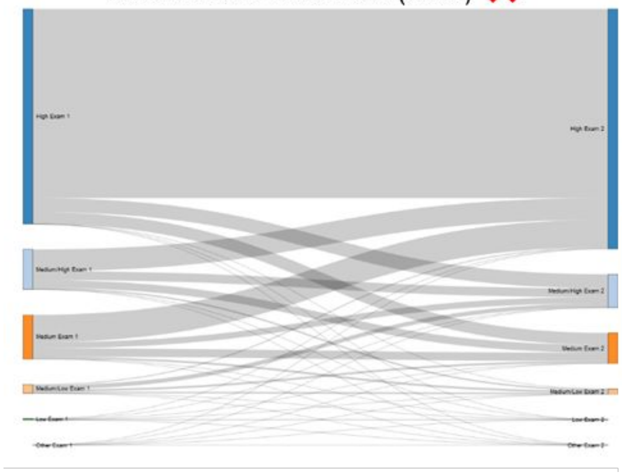
Stoichiometric Calculations (TG2) ✗



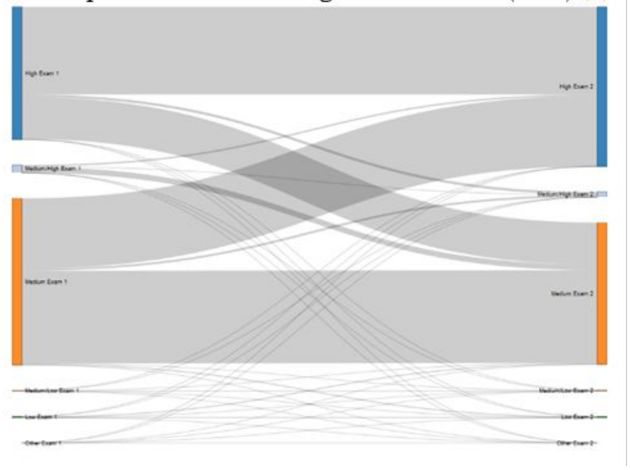
Limiting Reactant Calculations (TG2) ✗



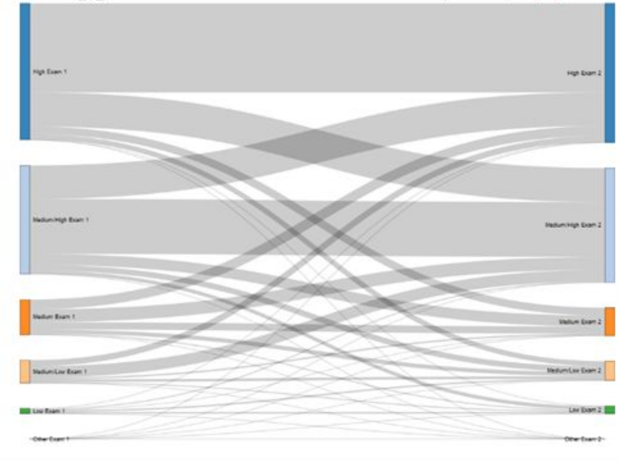
Chemical Formulas (TG3) ✗



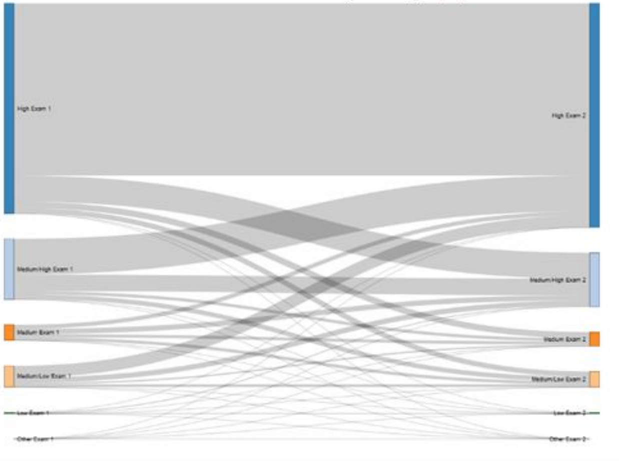
Conceptual Understanding of Molar Mass (TG3) ✗



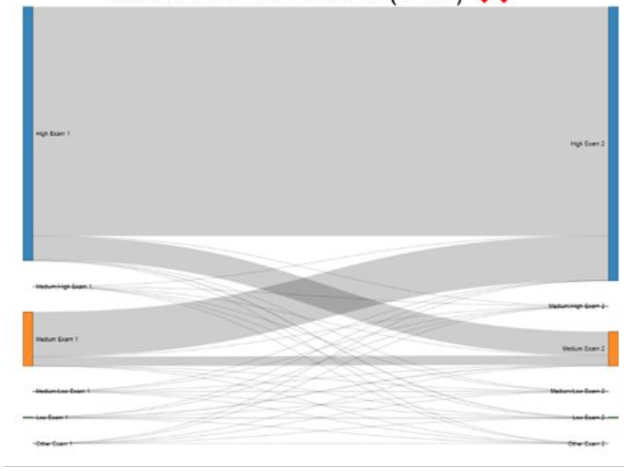
Application of Molar Mass (TG3) ✗



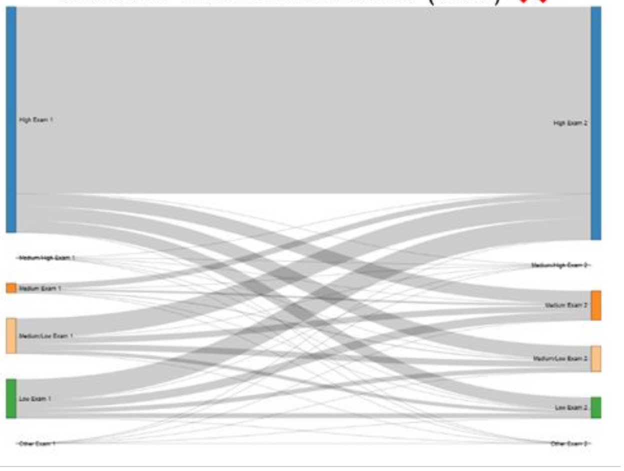
Mass Percent (TG3) ✗



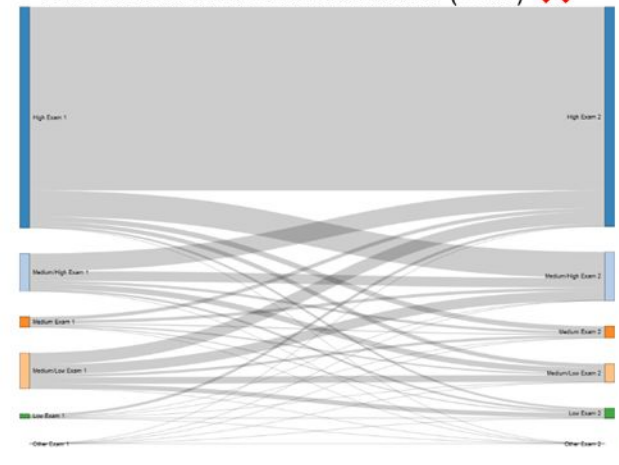
Mole to Mole Ratio (TG3) ✗



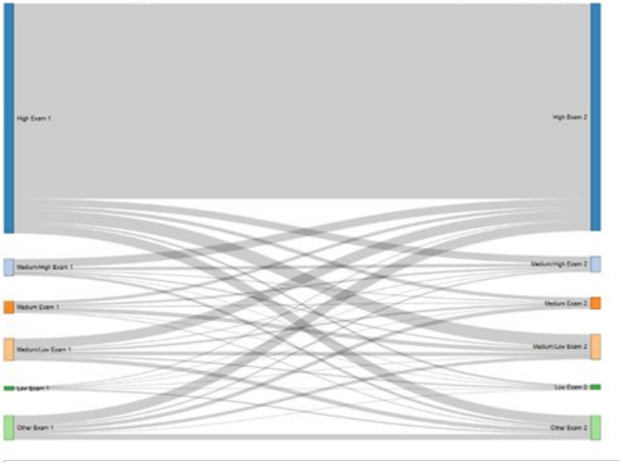
Mole to Mole Conversion (TG3) ✗

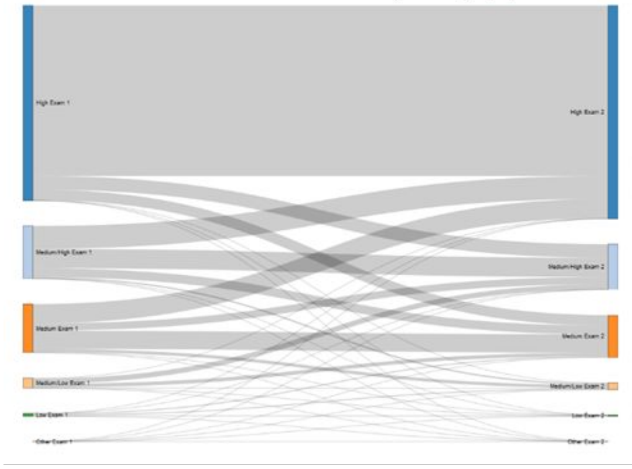


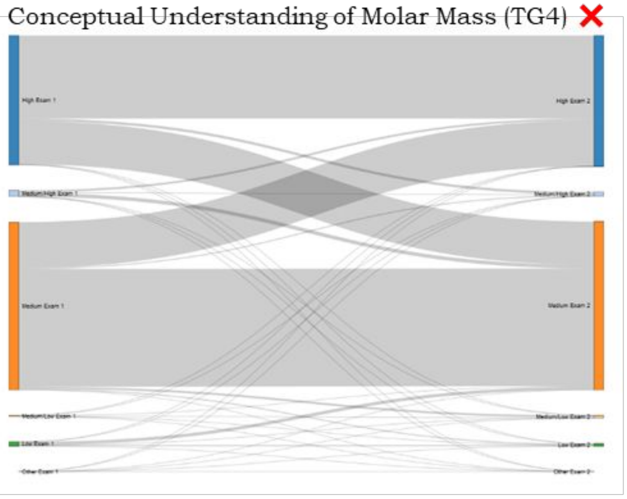
Stoichiometric Calculations (TG3) ✗

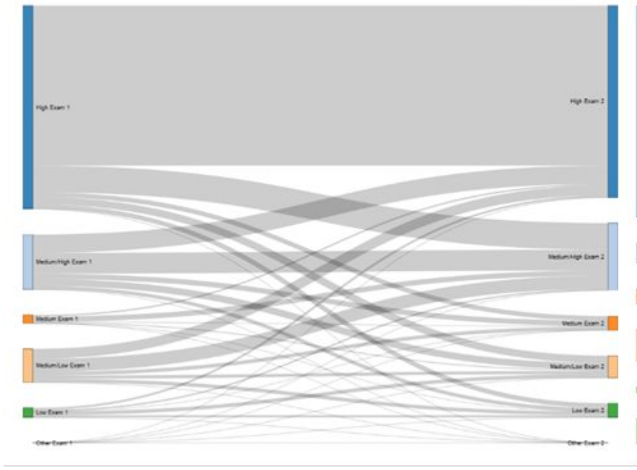


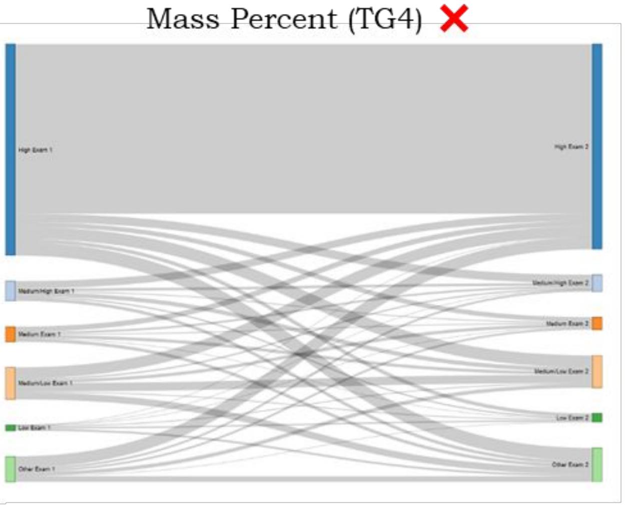
Limiting Reactant Calculations (TG3) ✗

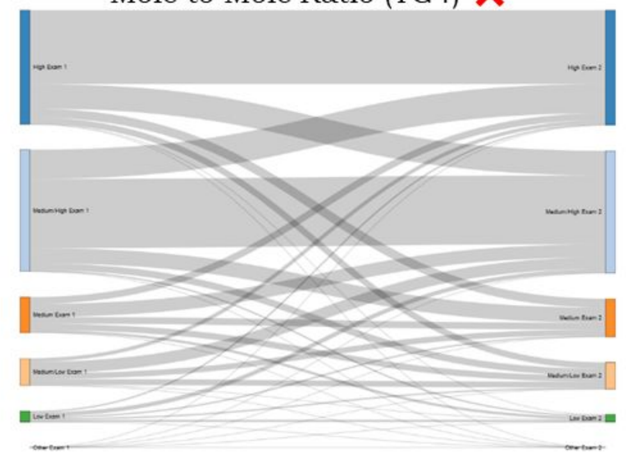


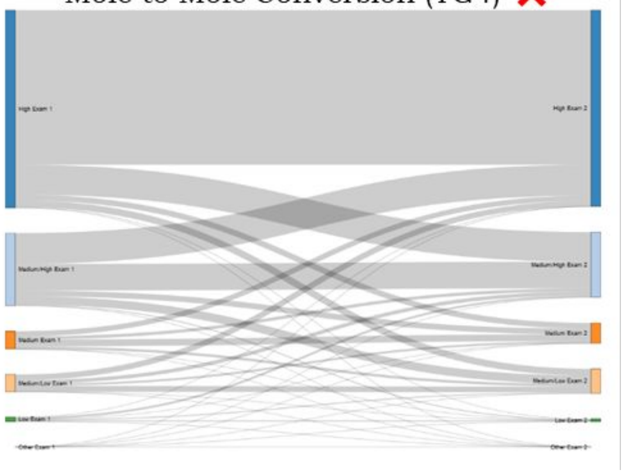

Chemical Formulas (TG4) ❌


Conceptual Understanding of Molar Mass (TG4) ❌


Application of Molar Mass (TG4) ❌


Mass Percent (TG4) ❌


Mole to Mole Ratio (TG4) ❌


Mole to Mole Conversion (TG4) ❌

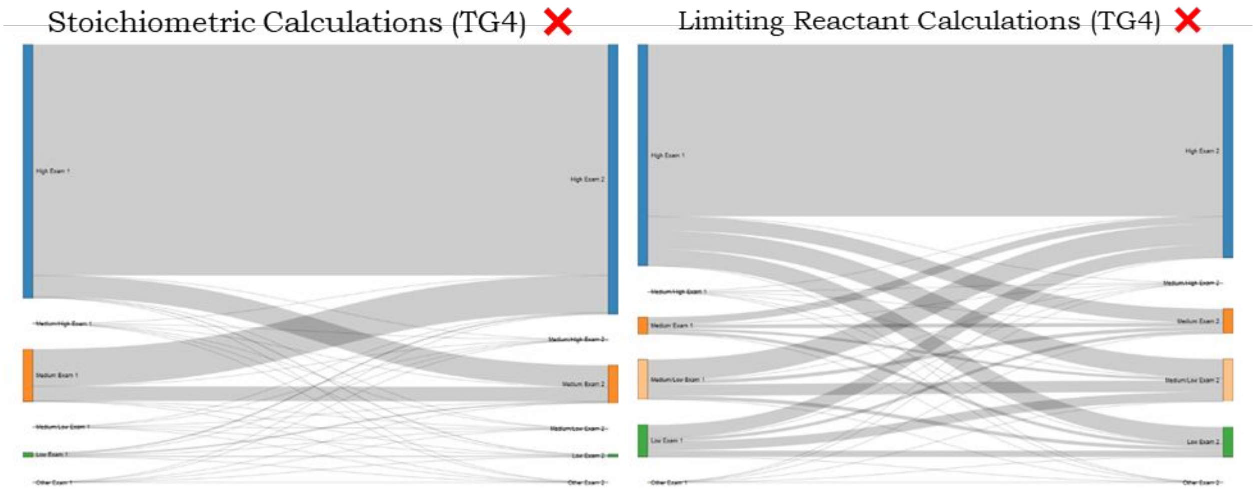


Fig. 12 Multimode ability and migration between weeks for each content area and Treatment Grouping (TG). Diagrams where dramatically more improvement was seen are marked with a "✓" where diagrams where improvement was canceled out by decline is marked with a "✗".

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).