Comparing Expert Predictions to Student Performance on Challenging Conceptual Questions: Towards an Adaptive Learning Module for Materials Science

Abstract

The emphasis on conceptual learning and the development of adaptive instructional design are both emerging areas in science and engineering education. Instructors are writing their own conceptual questions to promote active learning during class and utilizing pools of these questions in assessments. For adaptive assessment strategies, these questions need to be rated based on difficulty level (DL). Historically DL has been determined from the performance of a suitable number of students. The research study reported here investigates whether instructors can save time by predicting DL of newly made conceptual questions without the need for student data. In this paper, we report on the development of one component in an adaptive learning module for materials science – specifically on the topic of crystallography. The summative assessment element consists of five DL scales and 15 conceptual questions This adaptive assessment directs students based on their previous performances and the DL of the questions. Our five expert participants are faculty members who have taught the introductory Materials Science course multiple times. They provided predictions for how many students would answer each question correctly during a two-step process. First, predictions were made individually without an answer key. Second, experts had the opportunity to revise their predictions after being provided an answer key in a group discussion. We compared expert predictions with actual student performance using results from over 400 students spanning multiple courses and terms. We found no clear correlation between expert predictions of the DL and the measured DL from students. Some evidence shows that discussion during the second step made expert predictions closer to student performance. We suggest that, in determining the DL for conceptual questions, using predictions of the DL by experts who have taught the course is not a valid route. The findings in this paper can be applied to assessments in both in-person, hybrid, and online settings and is applicable to subject matter beyond materials science.

Introduction

In the 21st century, educators are responding to a rapidly changing context that includes shifts in the social circumstances of learners [1] and the technology tools available for instruction [2]–[5]. With increases in online, remote, and hybrid delivery, tools are needed that both support instructor-student interactions as well as accommodate growing class sizes and a more diverse student body. Including an adaptive instruction and feedback system in courses, for instance, would benefit not only students from receiving resources as and when required, but also benefit instructors to easily support the large group of diverse students [6]. To address these issues, we are developing a set of conceptually-based adaptive learning modules.

Our initial proof-of-concept centers on the topic of crystallography for an introductory course in Materials Science and Engineering as shown in Figure 1. The module contains videos with interactive questions, formative and summative assessments through conceptually-oriented multiple-choice questions, more detailed interactive instructional tools, and supplemental instruction for students who need it.

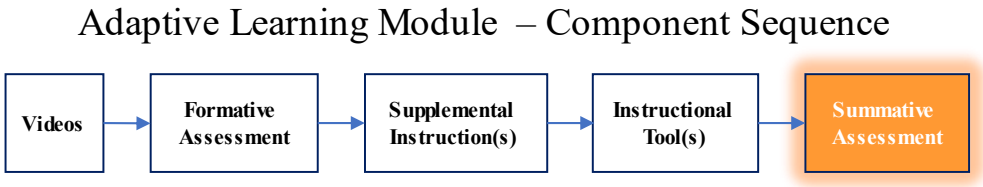


Figure 1 Simplified structure diagram of the adaptive learning module.

The research study reported in this paper is part of this larger project and focuses on the last highlighted adaptive component, the summative assessment. This element was designed as a progressive assessment process, where students who answered a question correctly would be directed to a more challenging conceptual question and those who did not select the correct answer would be steered to an easier question. Hence, we needed to rate the difficulty of the question – which became a challenging task in itself [7].

The fundamental question then became how good expert faculty who teach the course are at estimating conceptual question difficulty, and how do their estimates compare with student performance. Specifically, we address this research question:

> "*Do expert predictions of the difficulty level for challenging conceptual questions in crystallography match student performance?*"

While we studied this question in the context of developing the adaptive learning module, this fundamental question is of interest for educators seeking to use conceptually challenging questions in their instruction of assessment. If instructors can rely on their own predictions, they can provide students with questions of appropriate difficulty.

Background and theoretical framework

For Materials Science and Engineering, there is a growing focus on conceptual understanding. To this end, the use of concept questions has become common [8], [9]. However, when an instructor makes a new question, it is useful for them to have an idea of the question's DL. The common method to obtain a valid and reliable DL is by collecting sufficient student performance data and report an Item Difficulty Level [10], [11] and a Discrimination Index [10]. Automated methods such as the accumulative test by Sokolova *et al*. [12], the Question Classifier Engine by Narayanan *et al*. [13], or the algorithm based on a Monte-Carlo approach by Sud *et al*. [14] require actual student performance as input data. Therefore, it would be more efficient if we could rely on expert predictions of DL from. We term an *expert* as a person with deep content knowledge and considerable experience teaching that content in the classroom.

We have found only one other similar study. Pinter *et al.* [15] have compared student performance and their subjective estimations of the question difficulty with expert estimations. That study focused on basic programming knowledge and compared data among different years in bachelor's degree by using expert estimations as a standard. Indirect results from the study show some inconsistency between the expert estimations and student performance, however, no clear discussion or conclusions were made.

Elements of the adaptive learning module for crystallography

The adaptive learning module we are designing focuses on three core constructs in crystallography: crystal structures, atomic packing factor (APF), and theoretical density. We seek to develop student connected knowledge of atomic radius, atomic weight, lattice parameter, number of atoms per unit cell, coordination number, and basic crystal structure. Four types of basic crystal structures are used in the module: simple cubic, body-centered cubic, face-centered cubic, and hexagonal close-packed structures.

Figure 1 (in the Introduction section) presented the foundational elements of the adaptive learning system. Both the post-assessment logic of the formative assessment and the summative assessment are adaptive. In this paper, we focus only on the process of developing effective summative assessment (the highlighted component in Figure 1). As shown in Figure 2, the summative assessment is adaptive with possible paths students may take. The assessment in this module consists of three sets of questions. Each set represents a specific *construct* (crystal structure, APF, theoretical density) with five different questions. Each concept question in this assessment is a multiple-choice question. Each question is labeled with a different DL corresponding to the level of difficulty. Students start at the middle difficulty (DL3). A harder question is presented to the student based on a previous correct answer (indicated with "Y" as correct answer) or an easier question based on a previous incorrect answer (indicated with "N"). Thus, each student will answer three different questions per construct, but the questions will differ based on the demonstrated mastery. Results are recorded for further feedback on their overall performance.

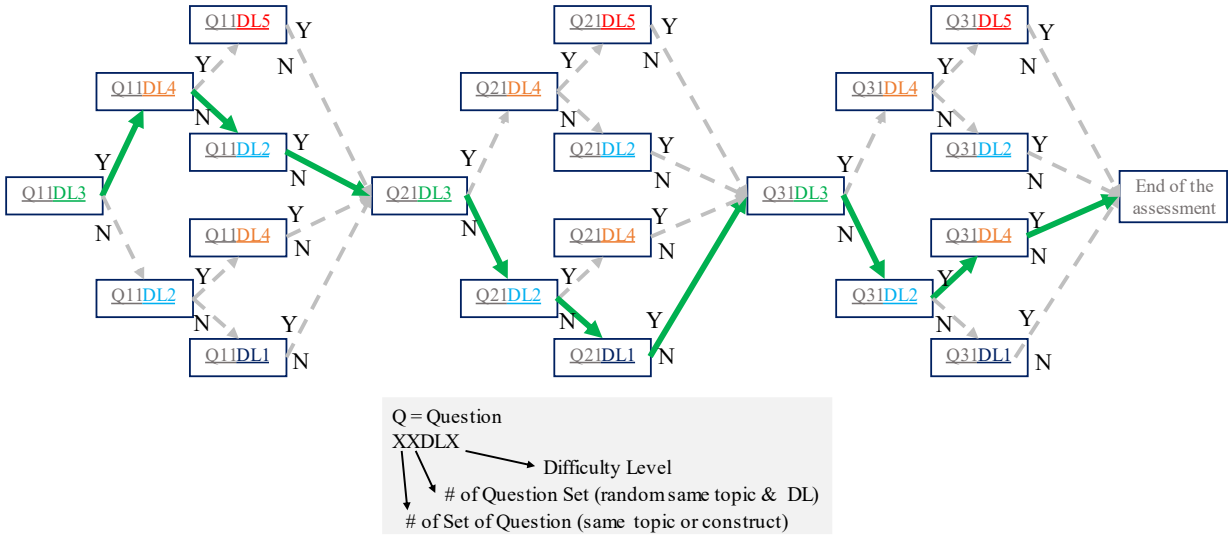


Figure 2 A diagram shows different possible paths in the summative assessment. Solid green arrows show a sample student path. Dashed arrows show other possible paths.

Methods and participants

We report results from creating the adaptive summative assessment. Using convenience sampling, we collected data from students enrolled in an Introductory Materials Science course

at a large, public institution as they were learning crystallography. Students in this class are mostly undergraduates in their junior or senior year pursuing a mechanical engineering degree. The Concept Warehouse (CW) [16] was used to assign the concept questions and collect student responses. All students were presented with a consent form when they first logged on to CW. For comparison with the expert data, we collected question responses from 439 students from 8 courses over 7 terms. These courses were delivered in either synchronously in-person, synchronously online, or asynchronously online formats. In all delivery modes, the same curriculum was used. The experts in this study were five professors who taught the Introductory Materials Science course in the past at this university. Their experiences varied from teaching three terms to about 40 terms, and the last term they taught the course varied from less than a year to over five years before this study.

The 15 conceptual questions studied are multiple choice questions with four answer choices, except two questions (ID 5247 and 5248), which have three and seven answers, respectively. After obtaining student results from the first term (121 students), all questions were revisited, and two questions were revised for future use based on their lack of clarity.

Measured and predicted DL refer to the actual student performance data and the expert predictions, respectively. For measured DL, a standard course grading scale was used as a criterion, as shown in Table 1. The measured DLs were obtained from actual student data using assumptions of a normal distribution curve with equal z-value bins. The predicted DLs were obtained from the experts who were asked to predict how they thought the majority of students would perform using the criteria in Table 1. The questions were provided in a random order but grouped together from the same construct. For the first round of prediction, all experts received the same no-answer-key package of these conceptual questions. They had time to work on these questions individually. After all data were collected, they joined a video conference meeting with the first author to look at the answer key, one question at a time. The experts had a chance to modify their previous predictions. During this conference, the experts could discuss the questions and provide suggestions about the questions and answers without revealing their initial predictions and whether they changed their predictions or not. Final predictions from each expert were collected at the end of the meeting.

Table 1 Compare the proposed DL with a standard course grading scale.

Difficulty Level (DL) (1 easiest – 5 hardest)	Standard course grading scale	
	Lower limit	Upper limit
1	90%	100%
2	80%	< 90%
3	70%	< 80%
4	60%	< 70%
5	0%	< 60%

In this paper, we report and compare expert predictions, both without and with an answer key. For easier presentation, expert predictions of the percentage of students that should get the

question correctly for all 15 conceptual summative questions (including two revised questions) were changed into DL values using the criteria from Table 1. We also report actual student performance for each course as well as the cumulative result of all 8 courses. We then compare expert final prediction with the cumulative result of the student performance.

Findings

Comparing initial and final expert predictions

The initial and final average predicted DL of experts are shown in Table 2, as well as standard deviations. The table also includes the number of changes in predictions after they saw the answer key for that question and the differences of the average predicted DL (final – initial). The negative values indicate that after they saw the correct answers, the experts mostly changed their prediction to easier (more students answering correctly). The last (right) column shows the record of discussions as experts did not discuss all questions.

Excluding questions before revision (ID 5256, 5261), the experts suggested small corrections to five of the 15 questions during the discussion session. Four of these questions (ID 5247, 5248, 5253, 5259) have figures showing multiple unit cells, so the experts suggest that adding an explanation to the question would make the question clearer. Two out of these four questions (ID 5253, 5259) refer to both the allotropic and polymorphic terms, and the experts suggested using only “allotropic”. Another suggestion was to add the term “vibrating surface” to question 5255, titled “box on a shaking platform.” In addition, one of the experts mentioned that questions with a lot of information, such as the shaking platform, are not normally used in their course.

From Table 2, we can see that the number of changes in predictions varied between zero and one, except for only one question (ID 5248) that has two changes. Nine (60%) out of these 15 questions have at least one change. Four out of the five questions that have discussions (except ID 5255) have one or two changes in predictions. Three (60%) out of five experts did not change any of their predictions for these 15 questions.

When we take a closer look at these changes from Table 2, five (55%) out of nine questions have higher standard deviation, compared to the initial round. Most of the changes of the average predicted DL evaluate the difficulty of the questions as easier than their initial prediction. Having the answer key with these conceptual questions might make some of these experts believe that the questions are not as hard as when they must figure out the answers by themselves during the first round.

Next, we compare these predictions with actual student performance. For further comparison, we present only the final prediction of the average predicted DL.

Table 2 The average predicted DL (initial and final), standard deviations, changes in predictions, the differences between initial and final predictions, and discussions from five experts for the summative assessments. (Note: DL1 = easiest to DL5 = hardest)

Construct	CW ID	Initial predicted DL		Final predicted DL		Differences in average predicted DL (final – initial)	Number of changes in prediction	Discussion (Y/N)
		Average	Std. Dev.	Average	Std. Dev.			
Crystal structure	5247	2.0	0.63	2.2	0.75	0.2	1	Y
	5248	3.0	1.10	3.0	0.63	0.0	2	Y
	5249	2.8	0.75	2.6	0.80	- 0.2	1	N
	5250	3.2	0.75	3.0	0.89	- 0.2	1	N
	5251	3.0	0.63	3.0	0.63	0.0	0	N
APF	5252	1.8	0.75	1.4	0.49	- 0.4	1	N
	5253	2.6	0.80	2.4	0.49	- 0.2	1	Y
	5254	2.4	1.02	2.2	0.75	- 0.2	1	N
	5255	3.4	1.20	3.4	1.20	0.0	0	Y
	5256*	3.0	0.63	3.2	0.40	0.2	1	N
	5374**	2.4	1.20	2.4	1.20	0.0	0	N
Theoretical density	5257	1.8	0.75	1.6	0.80	- 0.2	1	N
	5258	3.8	0.98	3.8	0.98	0.0	0	N
	5259	3.4	0.80	3.2	0.98	- 0.2	1	Y
	5260	2.4	0.49	2.4	0.49	0.0	0	N
	5261*	4.2	0.75	4.0	0.89	- 0.2	0	Y†
	5382**	3.8	0.98	3.8	0.98	0.0	0	N
						Total	10	Y = 5

* Questions before revision.
** Revised questions.
The orange and blue colors are presented for easier comparison in either increasing/decreasing in standard deviation and the differences in average DL.
† Discussion from the experts is the same reason why the question was revised.

Comparing student performance with expert predictions

Following the criteria from Table 1, measured DL for each summative question for 8 courses over 7 terms are shown in Table 3 along with the number of students. The correct percentages for each course are shown in the Appendix. There are a total of 439 responses for each of the 13 original summative assessment questions and 318 responses for the two revised questions. The measured DL from cumulative data (not averages from 8 courses) is used for the comparison with expert predictions. The last (right) column shows the average final predicted DL and the discussions data from Table 2 for easier comparison.

Table 3 The number of students and used data and the measured DL of the summative assessments for each of the 8 Introductory Materials Science courses (7 terms).

Courses (Sync. = synchronously, Async. = asynchronously)		Sync. In-person	Sync. Online	Async. Online #1	Async. Online #2	Async. Online #3	Async. Online #4	Async. Online #5	Async. Online #6	**Cumulative data**	**Experts**
Number enrolled		173	175	40	44	26	37	42	42	**579**	**5**
Number of (complete) responses		121	135	37	33	23	29	33	28	**439**	
Construct	CW ID	Measured DL (1 easiest – 5 hardest)								**Measured DL from cumulative data**	**Final Predicted DL and Discussion (Y)** (Table 2)
Crystal structure	5247	2	3	2	3	4	3	2	2	**3**	2.2 **Y**
	5248	3	2	1	2	2	3	1	2	**2**	3.0 **Y**
	5249	2	3	1	1	1	1	1	1	**2**	2.6
	5250	4	4	2	4	5	4	3	4	**4**	3.0
	5251	3	4	3	3	3	3	2	4	**3**	**3.0**
APF	5252	1	2	1	1	1	2	1	1	**1**	1.4
	5253	2	2	1	2	2	2	1	2	**2**	2.4 **Y**
	5254	4	5	3	4	2	5	3	3	**4**	2.2
	5255	5	5	5	4	5	5	5	5	**5**	**3.4 Y**
	5256*	5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	5*	3.2
	5374**	N/A	5	5	5	5	5	5	5	**5****	**2.4**
Theoretical density	5257	2	2	1	1	1	1	1	2	**2**	1.8
	5258	2	2	1	2	1	2	1	1	**2**	**3.8**
	5259	2	3	1	2	2	3	2	3	**2**	3.4 **Y**
	5260	5	5	5	5	4	5	5	5	**5**	**2.4**
	5261*	3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	3*	4.0 Y†
	5382**	N/A	4	1	1	2	2	1	1	**2****	**3.8**

* Questions before revision. Used only one term with 121 responses.

** Revised questions. Used for all other courses with 318 responses.

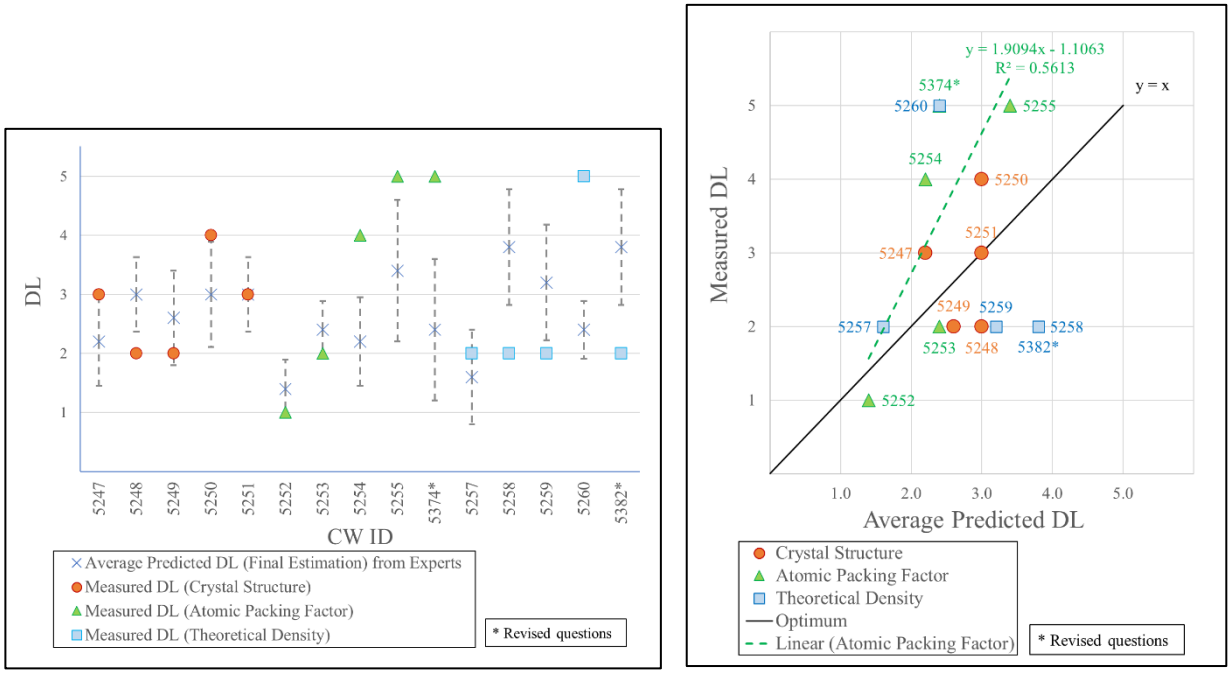
For the last (right) column, the orange and blue colors are presented for predictions with changes that are the deviated away or closer to the measured DL compared to the initial predictions, respectively. No change in prediction is represented with black color.

† Discussion from the experts is the same reason why the question was revised.

When comparing measured DL from cumulative data (Table 3) and predicted DL (Table 2), the average final predicted DL is closer to the measured DL than the initial one. Out of 10 changes in predictions, only three (in question ID 5250, 5254, and 5257, labeled the last right column in Table 3 by orange color) – one per construct – deviated away from the measured DL. In other words, seven (70%) of the experts' changes led to more accurate predictions of the actual student performance. In addition, four out of six questions that have discussions (indicated with "Y" and blue color) result in closer predictions.

The measured DL for each course (Table 3) shows fluctuations. Although using the same curriculum, other factors such as different settings (in person/synchronous online/asynchronous online), different instructors, student populations, and background in different terms might affect the variation between courses. This variation makes it harder for experts to predict with a small interval criterion (10%, Table 1).

To visualize this phenomenon, we plot between the measured DL from students and the (final) predicted DL from experts in Figure 3 (a), along with standard deviation error bars. The correlation plot of the same data is shown in Figure 3 (b) with a trend line only for the APF construct. The separated plots for each construct with standard deviation error bars and trend lines are shown in the Appendix.



(a) (b)

Figure 3 (a) A comparative plot and (b) a correlation plot between the measured DL and the average final predicted DL for the 15 summative assessment questions, separated by construct. The error bars on plot (a) are standard deviations for the average predicted DL. (Note: DL1 = easiest to DL5 = hardest)

From looking at the measured DL, questions in the APF construct covered the most DLs (four out of five). Measured DL for seven (47%) out of 15 questions lies within deviation ranges of

predicted DL with only one question (ID 5251) that matches with the average. Interestingly, this matched question had no discussion nor change in prediction by the experts (Table 3). As can be seen from both Figure 3 (a) and (b), there is no clear correlation between expert prediction and student actual performance - answering our *research question* ("*Do expert predictions of the DL for challenging conceptual questions in crystallography match student performance?*").

Conclusions

Identifying the difficulty level (DL) of conceptual questions is challenging. This project focused on the difficulty of new conceptual questions for crystallography to be used in the context of adaptive assessment. However, DL would also be useful for instructors to know if they plan to use this type of question for a classroom learning activity, or for an assessment of student understanding such as during course exams.

Five experts who have taught the course participated in our two-step prediction of the questions' DL – individually without an answer key followed by reviewing their predictions with an answer key in a group discussion setting. Actual student performance was collected for comparison from 439 students in 8 introductory Materials Science courses over 7 terms.

Data comparison between experts' initial and final predictions as well as number of changes in predictions were discussed. Comparing expert predictions of the DL with the measured DL from student performance shows no clear correlation, suggesting that it is not appropriate to depend only on the expert predictions. Discussion among the experts seemed to improve their predictions, but more study on the influence of discussion is warranted. In the context studied here, there was fluctuation among the student performance, even though the same curriculum was used across courses. We conclude that, for a limited number of experts, it is best to collect actual student performance instead of only using expert predictions to determine the level of difficulty for new conceptual questions, but that expert review can help refine the questions for clarity.

Acknowledgement

The authors acknowledge the support from the Division of Undergraduate Education, National Science Foundation (Grant #2135190), the California Education Learning Lab at the State of California, and the 2021 – 22 Ecampus Research Fellows Program at Oregon State University.

References

[1] B. J. Sottile, L. E. Cruz, Y.-A. lo Burleson, and K. McLain, "It's About Time: An Analysis of Student Activities Under Remote Learning," 2021. [Online]. Available: https://peer.asee.org/37412

[2] M. D. Koretsky and A. J. Magana, "Using technology to enhance learning and engagement in engineering," *Advances in Engineering Education*, vol. 7, no. 2, p. 53, 2019, [Online]. Available: https://eric.ed.gov/?id=EJ1220296

[3] M. Shuey, A. Akera, S. Appelhans, A. Cheville, T. de Pree, and S. Fatehiboroujeni, "Student Experience with COVID-19 and Online Learning: Impact of Faculty's Ability to

Successfully Navigate Technological Platforms for Remote Instruction," 2021. [Online]. Available: https://peer.asee.org/37742
[4] N. A. Libre, "Activating and Engaging Students in Online Asynchronous Classes," 2021. doi: 10.18260/1-2--36643.
[5] R. Subramanian, "Cutting-edge Tools & Technologies: Teaching Engineering Online," 2021. [Online]. Available: https://peer.asee.org/36291
[6] H. R. Weltman, V. Timchenko, H. E. Sofios, P. Ayres, and N. Marcus, "Evaluation of an adaptive tutorial supporting the teaching of mathematics," *European Journal of Engineering Education*, vol. 44, no. 5, pp. 787–804, 2019, [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/03043797.2018.1513993
[7] M. Raykova, H. Kostadinova, and G. Totkov, "Adaptive test system based on revised Bloom's taxonomy," in *ACM International Conference Proceeding Series*, 2011, pp. 504–509. doi: 10.1145/2023607.2023692.
[8] S. J. Krause, J. C. Decker, J. Niska, and T. Alford, "A Materials Concept Inventory for introductory materials engineering courses," in *National Educators' Workshop: Update 2002*, 2002, pp. 413–424.
[9] K. L. Kitto, "Perspectives from the Classroom - Developing Effective Concept Questions and Collaborative Learning for an Introductory Materials Engineering Course," in *36th ASEE/IEEE Frontiers in Education Conference*, 2006, pp. 1–5. doi: 10.1109/FIE.2006.322327.
[10] R. B. Livingston, V. L. Willson, and C. R. Reynolds, *Measurement and assessment in education*, 2nd ed. Upper Saddle River, N.J.: Upper Saddle River, N.J. : Pearson, 2009.
[11] R. Lourdusamy and P. Magendiran, "A systematic analysis of difficulty level of the question paper using student's marks: a case study," *International Journal of Information Technology*, vol. 13, no. 3, pp. 1127–1143, 2021, doi: 10.1007/s41870-020-00599-2.
[12] M. Sokolova and G. Totkov, "Accumulative question types in e-learning environment," in *ACM International Conference Proceeding Series*, 2007, pp. 1–6. doi: 10.1145/1330598.1330693.
[13] S. Narayanan, V. S. Kommuri, N. S. Subramanian, K. Bijlani, and N. C. Nair, "Unsupervised Learning of Question Difficulty Levels Using Assessment Responses BT - Computational Science and Its Applications – ICCSA 2017," in *Computational Science and Its Applications – ICCSA 2017*, 2017, pp. 543–552.
[14] P. Sud, M. West, and C. Zilles, "Reducing Difficulty Variance in Randomized Assessments," 2019. doi: 10.18260/1-2--33228.
[15] R. Pinter, S. M. Cisar, A. Kovari, L. Major, P. Clisar, and J. Katona, "Case Study: Students' Code-Tracing Skills and Calibration of Questions for Computer Adaptive Tests," *Applied sciences*, vol. 10, no. 20, p. 7044, 2020, doi: 10.3390/app10207044.
[16] M. D. Koretsky *et al.*, "The AIChE 'Concept Warehouse': A Web-Based Tool to Promote Concept-Based Instruction," *Advances in Engineering Education*, vol. 4, no. 1, 2014.

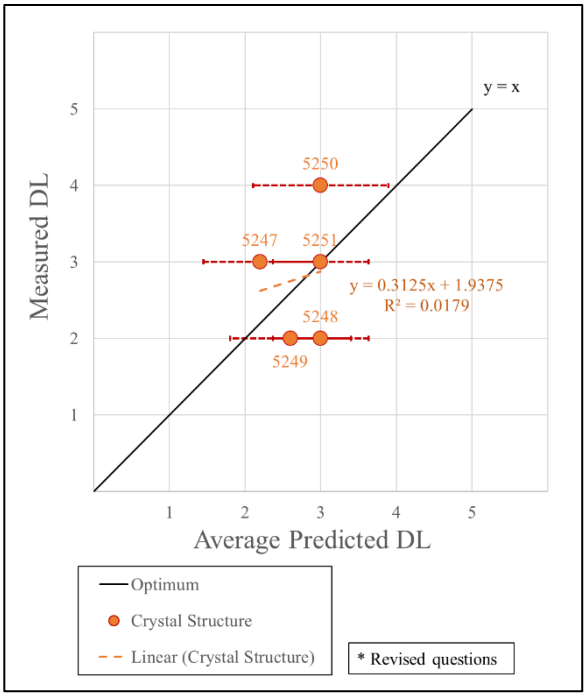
Appendix

Appendix Table 1 The number of students and used data and the correct percentage of the summative assessments for each of the 8 Introductory Materials Science courses (7 terms).

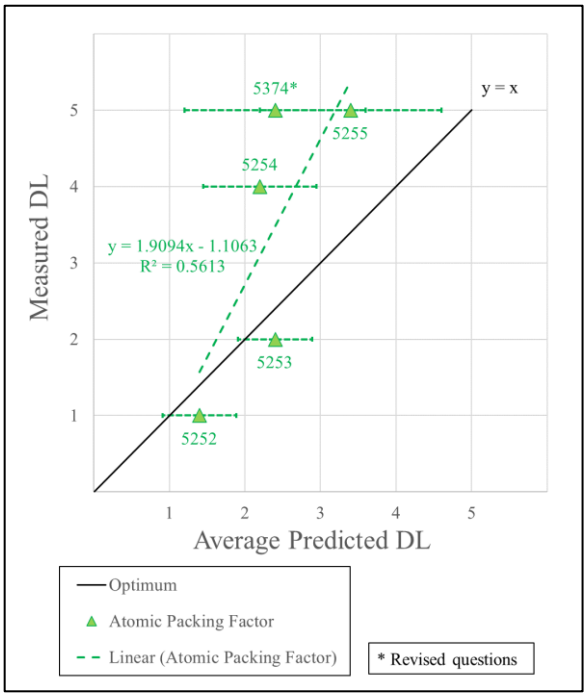
Courses (Sync. = synchronously, Async. = asynchronously)		Sync. In-person	Sync. Online	Async. Online #1	Async. Online #2	Async. Online #3	Async. Online #4	Async. Online #5	Async. Online #6	**Cumulative data**
Number enrolled		173	175	40	44	26	37	42	42	**579**
Number of (complete) responses		121	135	37	33	23	29	33	28	**439**
Construct	CW ID	Correct Percentage								
Crystal Structure	5247	80%	76%	81%	76%	70%	76%	88%	89%	**79%**
	5248	75%	81%	97%	85%	83%	72%	91%	86%	**82%**
	5249	86%	70%	97%	94%	91%	97%	91%	96%	**85%**
	5250	64%	60%	86%	67%	48%	66%	73%	61%	**65%**
	5251	76%	65%	73%	76%	78%	76%	82%	64%	**72%**
APF	5252	97%	88%	97%	91%	91%	90%	91%	96%	**92%**
	5253	89%	81%	97%	88%	87%	86%	94%	86%	**87%**
	5254	68%	56%	73%	70%	87%	59%	79%	75%	**67%**
	5255	37%	26%	27%	61%	52%	52%	48%	43%	**38%**
	5256*	21%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	21%*
	5374**	N/A	13%	30%	30%	30%	14%	27%	25%	**21%****
Theoretical Density	5257	88%	81%	95%	97%	91%	93%	100%	89%	**88%**
	5258	88%	80%	95%	88%	100%	86%	91%	96%	**87%**
	5259	83%	76%	92%	88%	83%	76%	85%	71%	**81%**
	5260	46%	30%	38%	48%	70%	31%	45%	57%	**42%**
	5261*	78%	N/A	N/A	N/A	N/A	N/A	N/A	N/A	78%*
	5382**	N/A	70%	92%	94%	83%	86%	91%	93%	**81%****

* Questions before revision. Used only one term with 121 responses.

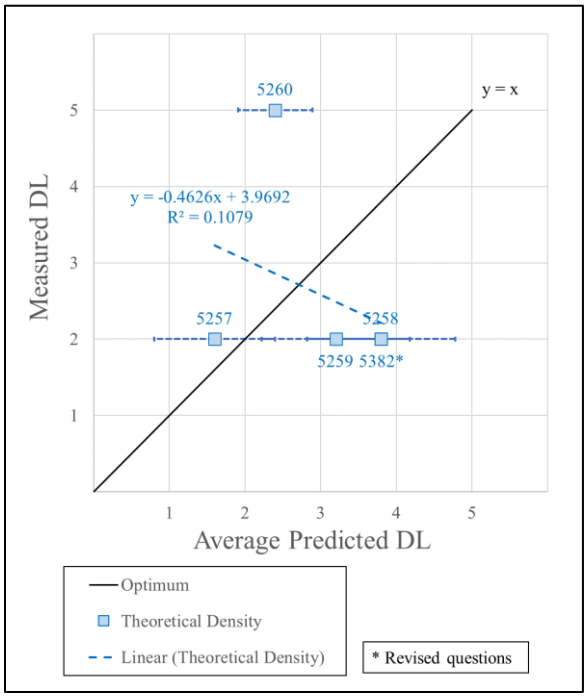
** Revised questions. Used for all other courses with 318 responses.



(a)



(b)



(c)

Appendix Figure 1 Correlation plots between the measured DL and the average final predicted DL for the 15 summative assessment questions with trend lines of (a) crystal structure construct, (b) atomic packing factor construct, and (c) theoretical density construct. The error bars are standard deviations for the average predicted DL. (Note: DL1 = easiest to DL5 = hardest)