Measuring Item Teaching Value in an Online Learning Environment

Jon Harmon
Macmillan Learning
211 E. 7th St.
Austin, TX 78701
+1 512-323-6565
jon.harmon@macmillan.com

Rasil Warnakulasooriya
Macmillan Learning
75 Arlington St.
Boston, MA 02116
+1 617-399-4531
rasilw@gmail.com

ABSTRACT
The Additive Factor Model (AFM) is a cognitive diagnostic model that can be used to predict student performance on items in a context that allows for student learning. Within AFM, *skills* have a learning rate, and student acquisition of a skill depends only on the number of opportunities a student has had to exercise that skill and the learning rate of that skill. Here we demonstrate an approach to measure the *teaching value* of individual *items* with respect to one another. The teaching values estimated through this approach may be useful for structuring intelligent tutoring systems and for content improvement.

Keywords
cognitive diagnostic models, item response theory, knowledge tracing, data mining, content pedagogy, content improvement, priming.

1. INTRODUCTION
Item Response Theory (IRT) models describe the performance of students with respect to a set of scored items (questions). As described by Sijtsma and Junker [5], most IRT models have three assumptions:

1. Local independence: Student performance on a given item does not depend on student performance on previous items.
2. Monotonicity: The probability of student success on an item increases when student ability improves.
3. Unidimensionality: Each student has the same ability for every item, and each item has the same difficulty for every student.

The simplest IRT equation, the 1-parameter logistic model or 1PL, can be expressed as

$$P(Y_{ij} = 1 | \alpha_i, \beta_j) = \frac{e^{\alpha_i - \beta_j}}{1 + e^{\alpha_i - \beta_j}} = f(\alpha_i - \beta_j)$$

where $f(x) = 1/(1 + e^{-x})$ (resulting in a probability in (0, 1)), Y_{ij} is the response of student i on item j (with 1 for correct, 0 for incorrect), α_i is the ability of student i, and β_j is the difficulty of item j.

Multidimensional IRT (MIRT) models relax the third assumption, decomposing student abilities and/or item difficulties into an array of abilities or difficulties. The Additive Factor Model (AFM) proposed by Cen et al [2] can be viewed as a MIRT model that also relaxes the local independence assumption, taking into account multiple exposures to a skill through a learning rate for that skill. The probability of student success is expressed as:

$$P(Y_{ij} = 1 | \alpha_i, \beta, \gamma) = f\left(\alpha_i + \sum_{k=1}^{K} \beta_k q_{jk} + \sum_{k=1}^{K} \gamma_k q_{jk} t_{ik}\right)$$

where $f(x)$ is again the logistic function $1/(1 + e^{-x})$, β_k is the difficulty (or easiness, with the sign reversed) of *skill* (rather than item) k, q_{jk} is a binary indicator that item j uses skill k, γ_k is the learning rate of skill k, t_{ik} is the number of exposures student i has had to skill k, and K is the total number of skills assessed. If each item addresses only one skill, the first sum reduces to B_j. If the learning rate is 0, the second sum reduces to 0, resulting in the 1PL equation.

In AFM, the likelihood of a student acquiring a skill is impacted only by the constant learning rate of that skill and the number of exposures that student has had to that skill. Here we seek to estimate a new quantity, the *teaching value* of an item with feedback. Exposure of a student to an item with a positive teaching value increases the probability that that student will answer a subsequent question correctly.

2. METHODOLOGY
Macmillan Learning's homework system examined in this study serves items (questions with automatic grading and feedback) to students, with the items grouped into assignments. The items investigated in this study were used across 570 institutions of higher learning in 4,704 courses. The order of the items in the assignments could be partially or fully randomized, and could be edited manually by individual instructors. As a result, a given item could be the first item in an assignment for some students, the second for others, etc., even within the same course.

We began with a dataset containing 240,990 unique students interacting with 7,257 unique general chemistry items, resulting in 29,005,495 student-item interactions. To simplify this proof of concept, we assume that all questions measured a single skill ("general chemistry knowledge"). To measure the impact of questions on one another, we define an *experience* as a set of one, two, or three items presented to a student in a particular order, starting with the first item in the assignment. For each experience,

Jon Harmon and Rasil Warnakulasooriya "Measuring Item Teaching Value in an Online Learning Environment" In: *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*, Collin F. Lynch, Agathe Merceron, Michel Desmarais, & Roger Nkambou (eds.) 2019, pp. 738 - 741

we scored the experience as *correct* (1) iff the student was scored as correct on their first attempt on the *last* item in that experience, and *incorrect* (0) if they were scored as incorrect on their first attempt on the *last* item in that experience, regardless of how they performed on any previous items. Using the dplyr package [7] for the R programming language [3], we filtered the full dataset to 32,199 unique students, 1,264 unique one-item experiences, 1,956 unique two-item experiences, and 1,567 unique three-item experiences, resulting in 3,240,791 student-experience interactions, using the following conditions:

1. The experience contained one, two, or three items.
2. Every item within each experience was attempted by at least 100 students as the *first* item in their assignment.
3. Every experience was attempted by at least 100 students.
4. Each student attempted at least 50 experiences.

By treating each experience as a single item, we were able to model the student abilities and difficulties of questions using a relatively simple IRT model. In this study we used the two-parameter logistic model (2PL) described by Birnbaum [1]. In the 2PL, items are allowed to have varying difficulties (β) and discrimination values (the ability of an item to differentiate between a low-skill student and a high-skill student, herein denoted a). The probability (P) of a student with ability θ answering a question correctly is given by:

$$ln\left(\frac{P}{1-P}\right) = a(\theta - \beta)$$

For the 2PL, it is assumed that the chance of a student *guessing* the correct answer is 0. The item and student parameters can be estimated using marginal maximum likelihood estimation (MML). Here we used the TAM package [4] to perform the estimations.

We fit the filtered student-experience interactions to a single two-parameter logistic model. To illustrate the approach in detail, our focus in this paper will be on the two-item experiences.

We define the *raw difficulty* (β_A) of a given item A as the modeled difficulty of item A when it is the *first* item attempted by students (i.e., when it is presented in a one-item experience). We define the *apparent difficulty* ($\beta_{A|B}$) of a given item A with respect to another item B as the modeled difficulty of item A when it appears *second* in a two-item experience, after item B. We calculated the *difficulty change* of a given item A after item B as the difference between the apparent difficulty and the raw difficulty for that item in that experience.

$$\Delta\beta_{A|B} = \beta_{A|B} - \beta_A$$

A *negative* difficulty change indicates that item A appears to be *easier* when it follows item B, and a *positive* difficulty change indicates that item A appears to be *harder* when it follows item B. For this study, we did not constrain the time between a student answering item A and that student answering item B.

3. RESULTS

Two "goodness of fit" statistics were used to evaluate the fit of each experience to the model: outfit and infit. Both measures are expected to have a value close to 1.0 for each experience if the model fits the data without overfitting or underfitting. 92% of experiences had outfit of 1.0 ± 0.05 (standard deviation), and 99% of experiences had infit of 1.0 ± 0.05 (standard deviation). We analyzed two-item experiences which met these criteria: 1) the experience had infit and outfit between 0.95 and 1.05, and 2) the second item of the experience had infit and outfit between 0.95 and 1.05 when it was in a one-item experience. Seventy six percent (1,490 of 1,956) of two-item experiences met these criteria.

The median difficulty change for these experiences was found to be −0.085, with the second quartile beginning at −0.42, and the third quartile ending at 0.27. We focused our analysis on experiences with difficulty changes more than 1.5 interquartile ranges above the third quartile or below the second quartile, and defined these difficulty changes as significant. Other difficulty changes may also be statistically significant, leading to net learning effects. However, the educational significance of such effects remains to be explored in future studies. We observed that 22 item pairs (1.5%) had a difficulty change more than 1.5 times the interquartile range below the first quartile, indicating a significant *decrease* in apparent difficulty (the items appeared to become significantly *easier*), as we expected would occur. Somewhat surprisingly, 17 item pairs (1.1%) had a difficulty change more than 1.5 times the interquartile range above the third quartile, indicating a significant *increase* in apparent difficulty (the items appeared to become significantly *harder*) (see Figure 1[1]). The remaining 1451 item pairs (97.4%) did not have a significant change in apparent difficulty as defined here. We examined specific cases of each type of change in apparent difficulty to attempt to identify the sources of the difficulty changes.

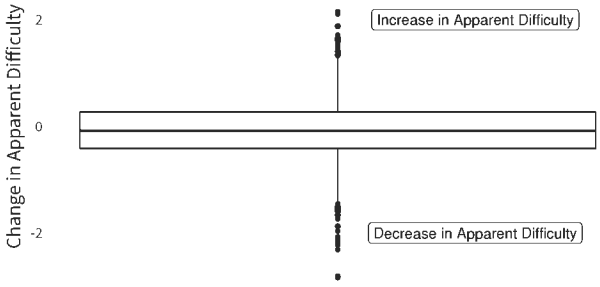

Figure 1. Most difficulty changes (the second and third quartile, shown in the box here) were near zero. However, some items showed a decrease in apparent difficulty (points at the bottom of the plot, lying more than 1.5 times the interquartile range below the second quartile), and some items showed an increase in apparent difficulty (points at the top of the plot, lying more than 1.5 times the interquartile range above the third quartile).

3.1 Decrease in Apparent Difficulty

Here we examine a specific case of difficulty changes in which the second item appears to be *easier* for students who have been primed by a specific preceding item. The item we designate "item A" in this study asks the student to

Give the conjugate acid for each compound below.

with three randomly selected bases such as HSO_4^-, CO_3^{2-}, and NH_3. The student must enter the conjugate acid for each base (in this example, H_2SO_4, $HCO3^-$, and NH_4^+). The item we designate "item B" is effectively the opposite question, in which the student is given three acids, and asked to enter the conjugate base. Even when the

[1] All figures were generated using the ggplot2 package for R [6].

randomly selected acids and bases do not line up from item B to item A for a given student, students perform significantly better when primed by item B before answering item A, or *vice versa.*

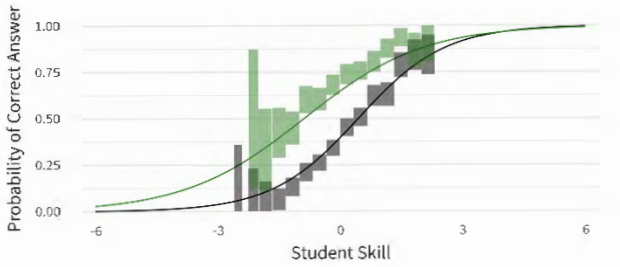

Figure 2. When a specific item A followed a specific item B (green), students performed better than when item A was the first question (gray); the curve shifted up and to the left, indicating that students with lower skill became more likely to answer the question correctly. The curves indicate the modeled difficulty and discrimination for each experience, while the boxes indicate the 95% confidence interval of actual student performance (% of students in a given modeled skill group who scored correct on the experience on their first try). See Figure 3 for the same items in the opposite order.

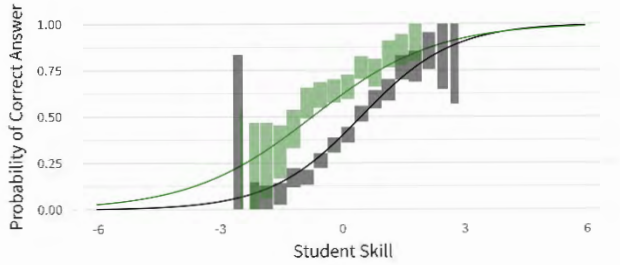

Figure 3. When item B followed item A (green), students performed better than when item B was the first question (gray). See Figure 2 for the same items in the opposite order.

In contrast, the item we designate "item C" is within the same general topic, asking the student to

> Label each reactant and product in this reaction as a Brønsted acid or base.

and showing the reaction between HCN and NH_2^- to produce CN^- and NH_3. Item C does *not* have a significant impact on student performance on item A (data was not available for student performance on item B after item C).

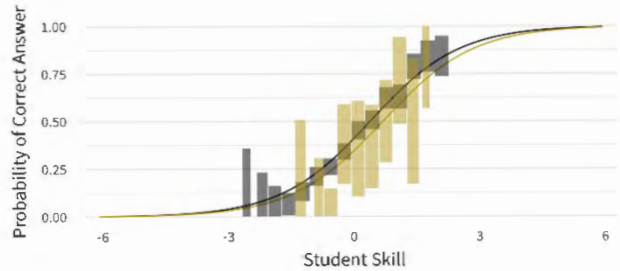

Figure 4. When item A from Figures 2 and 3 followed a third specific item C (yellow), there was no significant change from when item A was the first question (gray).

The raw difficulty for item A was found to be 0.45 ± 0.05 (95% confidence interval reported on all fit difficulties). When item A followed item B, the apparent difficulty was −0.9 ± 0.1, a difficulty change of −1.4 ± 0.1 (see Figure 2). Similarly, the raw difficulty for item B was 0.5 ± 0.1 and the apparent difficulty after item A was −0.7 ± 0.1, resulting in a difficulty change of −1.2 ± 0.1 (see Figure 3). In contrast, when item A followed item C, the apparent difficulty was not significantly different from the raw difficulty (0.7 ± 0.4 vs 0.45 ± 0.05) (see Figure 4).

3.2 Increase in Apparent Difficulty

Here we examine a specific case of difficulty changes in which the second item appears to be *harder* for students who have been primed by a specific preceding item. The item we designate "item D" asks,

> Parts per million (ppm) is a common way to express small concentrations of a solute in water. A sample of tap water that is 25 ppm Cl^- contains 25 grams of Cl^- for every 1,000,000 grams of water. Which units are numerically equal to ppm for dilute aqueous solutions?

Students are given 5 choices in a random order: "g/L", "cg/L", "mg/L", "µg/L", and "ng/L". The correct answer is "mg/L".

A possible explanation for this effect can be found by examining the terminology used in each item, and the particular choices that students selected. Item E asks students to

> Match each term with its definition or description.

with eight terms, including "parts per million." In the item we designate "item E," parts per million is defined as "micrograms of analyte per gram (or mL) of sample." In other words, Item D defines ppm in terms of L ("mg/L"), and item E defines ppm in terms of mL ("µg/mL"). Students who answer item E before item D are more likely to answer item D incorrectly, and they are more likely to do so by choosing the incorrect answer "µg/L" (see Figure 5). These results imply that students were led astray by the similarity of numerators of the units.

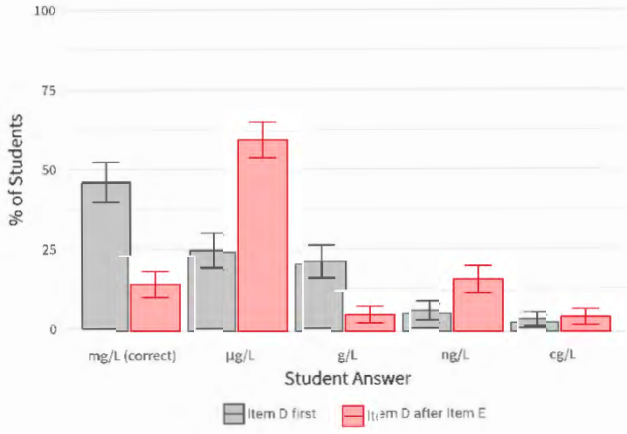

Figure 5. When item D followed item E (red), students chose "µg/L" more often than when item D was the first question (gray).

The raw difficulty of item D was 0.2 ± 0.3. When item D followed item E, the apparent difficulty was 1.9 ± 0.4, a difficulty change of 1.7 ± 0.4 (see Figure 6).

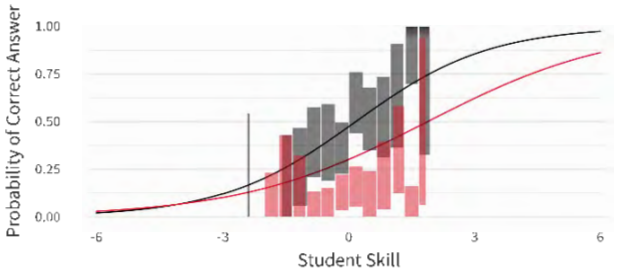

Figure 6. When item D followed item E (red), students answered incorrectly more often than when item D was the first question (gray); the curve shifted down and to the right, indicating that students required higher skill to have an equal probability of answering correctly.

4. CONCLUSION

We have shown that, adopting the methodology described above, we can measure the impact of individual questions with feedback on the difficulty of subsequent questions. This approach has implications for developing and revising pedagogically sound content. This approach could also influence sequencing of content to reinforce learning. Further research is required to attempt to generalize this difficulty change into a per-item "teaching value" parameter, but there appears to be evidence for this approach.

Items can have both positive and negative priming effects on the apparent difficulties of other items. Further research is required to determine the extent to which this effect increases or decreases with greater time in-between items, and whether intervening items have an impact on this effect.

5. ACKNOWLEDGEMENTS

JH would like to thank Jonathan Bratt for serving as a sounding board as this process was developed and for continued feedback throughout the process. JH and RW thank Adam Black for the opportunity to conduct this study.

6. REFERENCES

[1] Birnbaum, A. 1968. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical Theories of Mental Test Scores.* 397-472. Reading, MA: Addison-Wesley Publishing.

[2] Cen, H., Koedinger, K.R., and Junker, B., 2006. Learning Factors Analysis - A General Method for Cognitive Model Evaluation and Improvement. . In *Intelligent Tutoring Systems: 8th International Conference, ITS 2006, Jhongli, Taiwan, June 26-30, 2006. Proceedings,* M. Ikeda, K. D. Ashley, and T.-W. Chan, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 164–175.

[3] R Core Team. 2017. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

[4] Robitzsch, A., Kiefer, T., and Wu, M. 2018. *TAM: Test Analysis Modules.* https://CRAN.R-project.org/package=TAM.

[5] Sijtsma, J., and Junker, B.W., 2006. Item response theory: Past performance, present developments, and future expectations. *Behaviormetrika,* 33(1), 75-102.

[6] Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. http://ggplot2.org.

[7] Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2018. *Dplyr: A Grammar of Data Manipulation.* https://CRAN.R-project.org/package=dplyr.