Title

Outline:

**INTRODUCTION**

1. Why (frequent) open-book, low-stakes, online tests - flipped, follow up but all in spirit of learning by testing and JiTT, low-stakes online testing increasingly popular (RA
2. What is a difficult question? Low-stakes, open-book different from high-stakes testing, "Desirable difficulty" - difficulty depends on context, concept of “time well-spent”, why am I testing - performance vs mastery
3. Why read this? Here, we present a method to align easiness and answer time in order to assess and improve the quality of items in open-book quizzes

**METHODS**

* description of data (1st year med and dental students, n=170, number of tests, cell biology biochemistry class, MCQs)
* used learning catalytics to administer quiz (cite reference)
* point to scripts that analyze specific bits of data
* link to toy data set together with scripts on github
* IRB approval
* measures: IRT, difficulty….TBD
* explain outliers, give numbers (histogram) - median makes us more robust to outliers

**RESULTS**

1. Implementation of low-stake quizzes in an undergraduate medical classroom
2. Students are highly motivated and take quizzes seriously - the other paper
3. Low-stakes open-book quizzes show poor item discrimination (tied to low difficulty) - which is desirable!
4. Item response time can be used to gauge engagement (caution with “cognitive”)
5. Aligning median response time and easiness provides information about quiz items - 4 different quadrants
6. Instructors can use TEL plots (level as in Bloom) to assess and improve item quality (first vs second edition - emphasize concept of continuous quality improvement)
7. Availability of analysis tool

**INTRODUCTION**

Low-stakes open-book testing (often online) is a powerful educational tool as it allows students to monitor and assess their own learning at regular intervals, not just in the final exam. The concept of “readiness assurance tests or RATs” originates from team-based learning (CITE Michaelson’s paper), but can be used as a general teaching tool to ensure preparation for class. medical schools and other institutions are increasingly turning towards flipped and collaborative pedagogies, , frequent low-stakes testing proves useful in several ways: 1) A typical flipped classroom curriculum requires students to complete preparatory assignments before coming to class. Low-stakes open-book online testing can help assess students' preparedness and hold them accountable, e.g. by assigning a percentage of their grade to scores obtained on the RATs. 2) Results of RATs can be used to drive team discussions and to identify and address misconceptions in a Just in Time Teaching fashion (REF). 3) F Testing also promotes learning in itself and in fact is one of the most effective study strategies to enhance long-term retention, comprehension and transfer - this has been called the "test effect" (CITE review). Thus assigning low-stakes quizzes is not only a way to hold students accountable and inform teaching  but should directly enhance their learning. With online quiz tools readily available (CITE learning catalytics paper) it is easy to assign low stakes quizzes to help students prepare for class.

By definition, those low-stakes, open-book quizzes are very different from high-stakes tests. This raises the questions of whether instruments designed for high-stakes testing (such as classical test theory or item response theory) can be meaningfully applied to low-stakes testing. This is true in particular for the concept of difficulty. In high-stakes testing, the difficulty of an item can simply be defined as the proportion of test takers who give the correct answer. In low-stakes, open-book testing, and especially when it is used to assess students’ preparedness, it is actually desirable for this measure to be close to one (i.e. for most students to answer an item correctly). This does not mean that all questions on such an assessment should be easy, it just means that we need another way of measuring difficulty. A question can be “easy” in the sense that most students arrive at the correct answer, but still be “difficult” in the sense that getting to this answer requires sustained and prolonged cognitive engagement. This may require time, but it is time well spent, as far as the learning goals are concerned. Thus, one single measure of difficulty no longer seems enough to capture the complexity of open-book low-stakes testing and the demands it places on students.

Here, we present a method of aligning easiness, answer time and cognitive level (as rated by the instructor) in what we call TEL (Time-Easiness-Level) plots. Visualising those three dimensions of a given test item gives instructors quick visual feedback on the demands of a quiz question in terms of time taken, easiness and how this aligns with the instructor’s own assessment of the cognitive level of the item. This can help assess the quality of quiz items and edit them for improvement. The method is implemented in a set of R scripts available on github (CITE lanalytics github). It takes a spreadsheet of quiz question answers and time stamps (for each student and each question) as an input, and can be used with data from any quiz taking software that records this type of information. We describe the use of TEL plots and of the *lanalytics* software in a first-year medical course and the results obtained over two years.

**METHODS**

**Setting and population**

This project was embedded in a first year cellular and molecular biology course at Harvard Medical School in both 2013 and 2014. Class size in both years was 170 students, including both medical and dental students. Each year, 29 low-stakes, open-book quizzes were made available to the students. Completion of a quiz carried a small amount of credit, provided a student had more than half of the answers correct. Compliance was very high, with most students completing and passing every quiz. Quiz questions were mostly multiple-choice, although some other question formats (many-choice, short answer etc.) were sometimes included. The analysis presented here is limited to multiple-choice items. Quizzes were administered using the Learning Catalytics software [CITE Learning Catalytics paper].

**Ethics**

The study is part of a larger study that received approval from the Institutional Review Board at Harvard Medical School.

**Data analysis**

Results (student identifiers, item answers, time stamps and correct/incorrect information) were exported in spreadsheet format. The first step in the data analysis pipeline was to run the data through a script that automatically assigned random four-digit IDs to student. The map between those auto-generated IDs and student identifiers was given to course faculty, but was not shared with the rest of the project team, and all further analysis was done on de-identified data.

All analysis scripts were written in R and are available on github (<https://github.com/MelanieIStefan/lanalytics>) under a GNU General Public License. The README.md file contains information about specific scripts and the order of dependencies.

**Measures and constructs**

measures: IRT, difficulty….TBD

* explain outliers, give numbers (histogram) - median makes us more robust to outliers

**RESULTS**

**Implementation of low-stakes quizzes in an undergraduate medical classroom**

In a class on molecular and cell biology for first year medical students we created a series of 29 low-stakes, open-book, online quizzes to promote preparation for small group PBL sessions (“tutorials”).  In order to prepare for tutorial students were asked to work through a set of learning objectives covered in 1-2 hours of lecture each day. Quizzes were made available to students after lecture and had to be completed before the tutorial sessions, often on the next day. The quizzes were designed to cover every learning objective with an average of 10 questions, mostly multiple choice (Table 1). If a student completed the quiz on time with more than 50% correct they were awarded a point towards their grade, with all 29 quizzes accounting for 20% of their final grade (the rest of the grade coming from two exams). Table 1 summarizes basic properties of these quizzes from 2 consecutive years.

Table 1: Comparison of basic quiz properties before and after revision using easiness and item response time (IRT)

|  |  |  |
| --- | --- | --- |
|  | 1st edition | 2nd edition |
| Quiz properties | | |
| Total number of items | 284 | 288 |
| Total number of quizzes | 29 | 29 |
| Ave # items per quiz | 9.8 | 9.9 |
| Ave quiz score (stdev) | 82.5 (15.8) | 84.1 (14.7) |
| Median quiz time (min) | 20.5 min | 34.7 min |
| Item psychometrics | | |
| Ave item difficulty (stdev) | 0.83 (±0.15) | 0.84 (±0.15) |
| Ave discrimination (stdev) | 0.15 (±0.11) | 0.15 (±0.11) |
| Median IRT (interquartile range) | 89s (158s) | 62s (155s) |
| Item format | | |
| % items multiple choice | 90.8 | 99.3 |
| % items short answer | 5.6 | 0.3 |
| % items other (e.g. ranking, sketching) | 3.5 | 0.3 |

**Low-stakes open-book quizzes show poor item discrimination**

Students completed the quizzes diligently and generally reported them as useful or very useful in the course evaluation (data not shown). In line with the open-book nature, quiz items on average were answered correctly by > 80% of the class and consequently had very low discrimination (Table 1; REF). This raised concerns among faculty about the quality of the quiz items, whether they were too easy and in fact useful for learning. However, the quizzes had been very well received and upon reflection we realized that in an open-book test designed to promote preparation for class, it was actually our goal that students succeeded at solving most items correctly. Thus the high average scores indicated that the quizzes were of some utility to the students, but we were still lacking a way to assess item quality and revise content meaningfully for the upcoming course.

**Item response time can be used to gauge students’ engagement**

In thinking about the process students would use to engage with the quizzes it occurred to us that the time it took students to answer a given question might enable us to understand, if a given item engaged the students appropriately to its content. Since the platform we used to administer these quizzes recorded a time stamp each time a students answered a question we indeed were able to calculate item response times (IRT) for each item.

Figure 1 - histograms of IRT plots of individual items, binned at 20 seconds or whatever looks good, use shades of grey to indicate correct vs incorrect answers. A) easy and fast, B) difficult and long OR examples of all 4 quadrants

Figure 1 shows two typical example of item response time. While the large majority of students answered around 1-2 minutes, depending on the difficulty of the item, some students took more time. Since these quizzes were open-book and could be paused at any time, occasionally students would start a quiz and then take a break leading to very long item response time. The overall frequency of these events was very low [% of all IRT longer than 20 minutes? OR present histogram]. Throughout this paper we will report the median IRT that is not sensitive to these outliers.

[Are long IRTs associated with a higher likelihood of getting the answer correct? Start with violin plots.]

The first time we included the quizzes in the course (1st edition) the median time students worked on each question was about 1.5 minutes. The median time students needed to complete a quiz was about 20 minutes (Table 1). Thus over the time of the course, students invested about 10 hours  study time total into this resource and the quizzes were met with very positive ratings in the end of course evaluations (data not shown). In the following year, the median item response time dropped to 1 minute [run unpaired t-test to prove significance] while the median time spent on a quiz increased by 75% to 35 minutes [run unpaired t-test to prove significance].

* Did we include the SA question in the total quiz time calculation?
* Where do students spent the additional time? Do they invest it in more difficult items?

**Aligning median response time and easiness provides information about quiz items**

Interestingy the effect of content matter of the item (factual vs pure comprehension or applied comprehension) on IRT and difficulty vaired quite a bit, in particular if only looking at a single quiz as the example in Figure 2. When analyzing all quiz item, it turned out that While factual items that tested terminology or very easy concepts tended to populate the upper left qudrane of fast and easy items [ANOVA] and were answered faster and more likely correctly byt the students, however no significant difference was found between items that tested comprehension

[try linear regression for F/C/A items separtely)

Figure 3

Easy and fast items: As you can see, questions that tested factual knowledge (Questions 2, 3 and 9) clustered in the lower left quadrant (fast and easy) with would be expected as students will either remember it or not or quickly can look it up min their notes using the text search function. This was the case globally, Table 3 [ANOVA of item classification vs easiness and IRT] - factual items tended to take less time to answer and overall students were more likely to get them correct, though it may not always be the case as demonstrated by question 11 in quiz B, suggesting that the related fact might not have been played out as clearly in the preparatory material.

Difficult and fast items: If students answer a question rapidly and incorrectly the assumption lays near that either the question or the resources contain misinformation or that the students hold a true misconception. No such example is included in the examples here and we have found this to be a rare event related to mistakes in the questions [Figure 2C. Confirm that this is correct by looking up the specific examples].

Easy and slow items: Item 5 quiz 8

Difficult and slow: Item 8/9/2 quiz 8 and items 4/10/7 quiz 9 - most

In general, 75% correct seems a reasonable threshold to ensure the students come to class prepared and ready work on novel assignments. In particular in formats that are collaborative like PBL or CBCL it is important that the team members similary prepared. I minute seems to be the time students need answer facitl questions that require less thinking, either from memory or by looking it up quickly. Tifgure 2 sumarizes how we thinkg IRT/easiness plots can be used to evaluate open-book online quiz item and allow faculty to approximate the utility of the quizzes for the students; learnin

g. The goal is to work collaborativley and while this will allow students to address some of their questios as part of the course work, if students came . Item 2 on the other hand the The stem of this item was revised to be less confusing and in the next year, about 75R of the students was able to solve this item at the first attempt. In general we beiliee that 75% is a good threshold to judge if an item Also, about a quarter of the students does not get this questions right at the first attempt and then might choose to restudy the respective part of the material. Items 8 and 9 did not take as long as item 10 and the class was mich less succesful in solving these two problems. The other pattern  many questions clustered in the lower right quadrant of the plot, meaning that each item cost students considerable time (> 1 minute) and that despite that investment of time more than a quarter of the class was still unsuccessful at solving them correctly. Upon checkin these and the related preparotory mateiral we found that

**Instructors can use TEL plots to improve quiz quality**

Figure 2 illustrates an example of how the revision process works on one quiz as example. Each quiz item is represented as a dot in an X by Y plot of median IRT vs. easiness (also called difficulty: the ratio of incorrect/correct answers). The color coding indicates the nature of the content tested by this item. Items that test factual knowledge or terminology are purple. Items that test understanding of a concept are orange, and items that test comprehension of a concept by asking students to apply it to new information are green. It is important to emphasize this classification is subjective and reflects the instructor’s opinion in context of the study resources provided to the students.  The dark lines divide the grit at 1 minute and 0.75 easiness (or 75% correct answers) into empirical quadrants the nature of which will be discussed below.

Figure 2 - note: need new image of quiz 8 from 2014 since item 9 does is not displayed (factual item), also for this particular item include item 3 although not MC

In the first edition of this particular quiz, question 5 took students very long although the concept tested by this item was deemed quite easy by the instructor. Closer examination revealed that one distractor introduced unnecessary confusion. The distractor was revised accordingly and in the next edition the item behaved fast and easy. It is important to note that without IRT, item 5 would not have stood out in need of revision.

Items 3 and 4 are not displayed in Figure 2 since in the first edition of this quiz a sketch format was used that proofed hard to grade and was replaced with multiple choice question in the 2nd edition. In general, for the 2nd edition we converted all question to multiple choice (Table 1) since other formats tended like sketching or matching were hard to grade or more easily misunderstood (data not shown).

Questions 10 represents an item that tested comprehension of one of the more challenging concepts in this unit. Accordingly it took students 3-4 times as long as the easy items. Yet, given that about 75% of the class succeeded in solving this questions as this seemed a desirable level of difficulty and the item was not edited. This is in contrast to item 8, which tested comprehension of a  concept less challenging than item 10 but almost half of the class failed this item. The more rapid IRT of this item compared to item 10 indicated that some students might have been unaware of the difficulty of this question. Upon close examination it seemed that the stem of the item was worded confusingly and might have mislead students to pick a certain distractor. Revision resolved this problem in the 2nd edition. Similarly, in item 9, just removing the distractor “all of the above” many more students were able to solve this item correctly.

Another interesting example is item 2, the most difficult item of all. Further examination revealed that item 2 and 10 pertained to the same challenging but core concept of the unit. Analysis of the related course resources revealed that the content was explained poorly and although revision of the materials might have contributed to the slightly shorter IRT of time 10 in the 2nd edition, the continued poor performance of item 2 clearly indicates further need of improvement. These examples illustrate how by triaging item difficulty and item response time combined with the intricate knowledge of the course material of the instructor can become a highly valuable and nuanced process that allows unbiased and unobtrusive insights into the learner experience of the quiz and the associated resources.

In gneral, items that were not edited and behaved very similarly in both years. [POSSIBLY ADD GENERAL SENTENCE HERE ON “ALL/NONE OF THE ABOVE” SCENARIOS].

[ADD GENERAL SENTENCE ON ITEMS THAT WERE UNCHANGED HERE].

**Availability of analysis tool**

The scripts used for data analysis are available on github (https://github.com/MelanieIStefan/lanalytics) under a GNU General Public License.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**DISCUSSION**

Testing has many benefits -  testing is known to promote learning by improving retentions as well as transfer of material to new content. Testing gives students instant feedback and allows them to adjust their preparation for class. Testing allows faculty to know which areas their class is mastering and where they benefit from further instruction. Yet, testing also introduces stress and students often perceive testing not as tool for their learning but solely as means to prepare for an exam.

t is well known that students are motivated by high-stakes assessments and will study long hours (“cram”) before exams but are much less likely to study on a regular basis, despite abundant evidence that distribute practice (such as preparing for class regularly) is much more beneficial for learning.

Include reference to Kornell and Mjoerk, 2007 that shows that students test to find out how well they have learned but only few students test FOR learning.(Psych Bulletin review, 6:219\_

* IRT is well known tool to distinguish motivational state of students and recently also has been used to assess utility of questions in class (Mazur paper)

**APPENDIX**

[CREATE APPENDIX FOR QUIZ 8 WITH VERSIONS FROM 2013 AND 2014].