How does the AMBOSS platform adapt to individual learner performance?

A deep dive into the data science and methodology behind longitudinal learner modeling.

Executive Summary

AMBOSS has developed a psychometric system that profiles a student's knowledge and delivers longitudinal insight into their performance. The system applies a knowledge tracing model to continuously estimate topic mastery across medical subjects and clinical competencies.

This model powers personalized practice sessions on the AMBOSS platform. Since the launch of Adaptive Qbank sessions in April 2026, over 36,000 students have used the feature to efficiently focus practice on areas of weakness. The AMBOSS platform also leverages the learner ability estimates to provide students with advanced insights on exam readiness.

In this white paper, we present the methodology underlying the AMBOSS platform's learner model. We also validate the accuracy of the learner ability estimates by comparing AMBOSS exam score estimates against actual learner performance on the USMLE Step 2 Clinical Knowledge (CK) exam.

Introduction

Question banks are central to high-stakes licensing exam preparation and daily study for medical students.[1] However, traditional question bank performance dashboards primarily summarize performance on previously completed questions within a study objective rather than estimate a learner's current knowledge. Consequently, these metrics depend on which questions a learner has completed and how they were selected, limiting their ability to provide a stable measure of exam readiness or carry forward evidence of learner knowledge across study objectives. And while most question banks surface weaknesses, they do so only at a macro level, providing limited guidance on what learners should study next.

The AMBOSS system addresses these limitations by applying the principles of precision medical education, which emphasizes the use of continuous learner data to provide personalized feedback and timely interventions.[2] Drawing on decades of research in knowledge tracing, AMBOSS developed a learner model that continuously estimates learner knowledge over time and serves as the foundation for Adaptive Qbank sessions.

The result is a more informative approach to question bank use: one that can estimate granular knowledge more reliably across time and study objectives, ensure learners work on exam relevant weaknesses earlier, and generate advanced insights into their exam readiness.

Method

Knowledge tracing broadly refers to estimating a learner's latent knowledge from sequences of observed responses over time.[3] Many knowledge tracing models exist, including Bayesian, logistic, and neural approaches.[4]

Rasch-based Learner Model

The AMBOSS platform implements a knowledge tracing system based on item response theory (IRT), a family of probabilistic models that relate a learner's latent ability to the likelihood of answering questions (i.e. items) correctly.[5]

The central challenge in learner modeling is that not all correct or incorrect answers provide the same information. Consider two learners. Alice is a third-year medical student deep into her dedicated study period for Step 2 CK who has completed thousands of practice questions and consistently performs well across a broad range of topics. Bob is a first-year medical student who has answered far fewer questions and is still building foundational medical knowledge.

Now suppose both learners answer the same question correctly. That single response should not necessarily be interpreted in the same way for both learners. Similarly, answering a particularly challenging question correctly may provide different information than answering a relatively easy question correctly.

The **Rasch model**, a variant of IRT, addresses this problem by modeling the probability of a correct response as a function of both learner ability and question difficulty (see Appendix for details).[6] For a given learner, more difficult questions are less likely to be answered correctly, while for a given question, learners with higher ability are more likely to answer correctly. This probabilistic relationship allows the Rasch model to predict a learner's success on individual items.

Figure 1 illustrates this probabilistic relationship between learner ability and question difficulty by plotting Item Characteristic Curves for hypothetical hard and easy questions. As the learner's ability increases, the probability of answering either question correctly also increases. The curve for the more difficult question is shifted to the right, indicating that a higher level of ability is required to achieve the same probability of success.

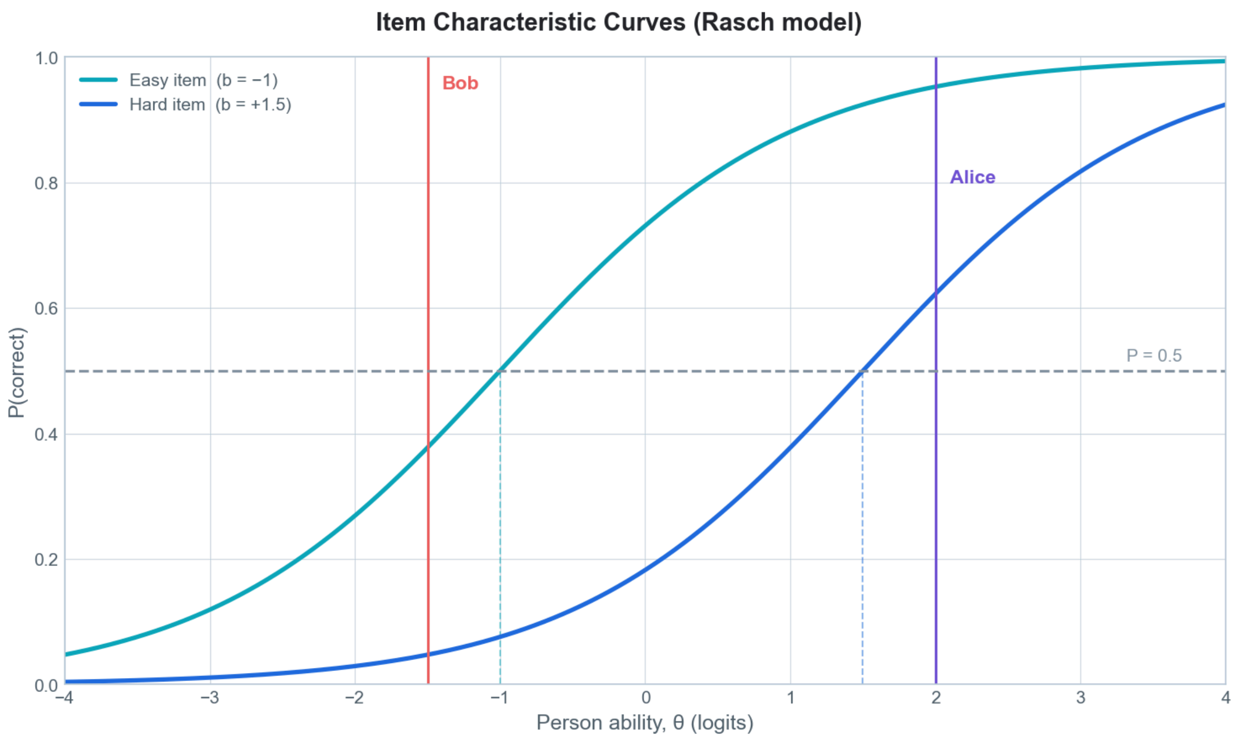

*Figure 1. **Item Characteristic Curves**. The y-axis represents the probability of a correct response to a question. The x-axis represents the student's latent ability. The point where the probability of a correct response is equal to 50% (y=0.5) is the difficulty level of the question on the x-axis.*

Many latent skills

The Rasch model provides a useful framework for estimating overall ability, commonly denoted by the parameter theta (see Appendix for details). However, a single ability estimate is insufficient for describing medical knowledge, which spans many domains and competencies.

The AMBOSS platform breaks down medical knowledge into skills across multiple dimensions, including discipline, body system, clinical task, and medical topic. Each question in the AMBOSS question bank is tagged with these skills. For the current version of the learner model, the following two skill dimensions are most relevant:

1. **competency-level skills**, which categorize knowledge by clinical tasks such as formulating diagnoses and planning treatment.
2. **topic-level skills,** which represent knowledge of medical topics such as "Pneumonia", "Myocardial Infarction" and "Arrhythmias".

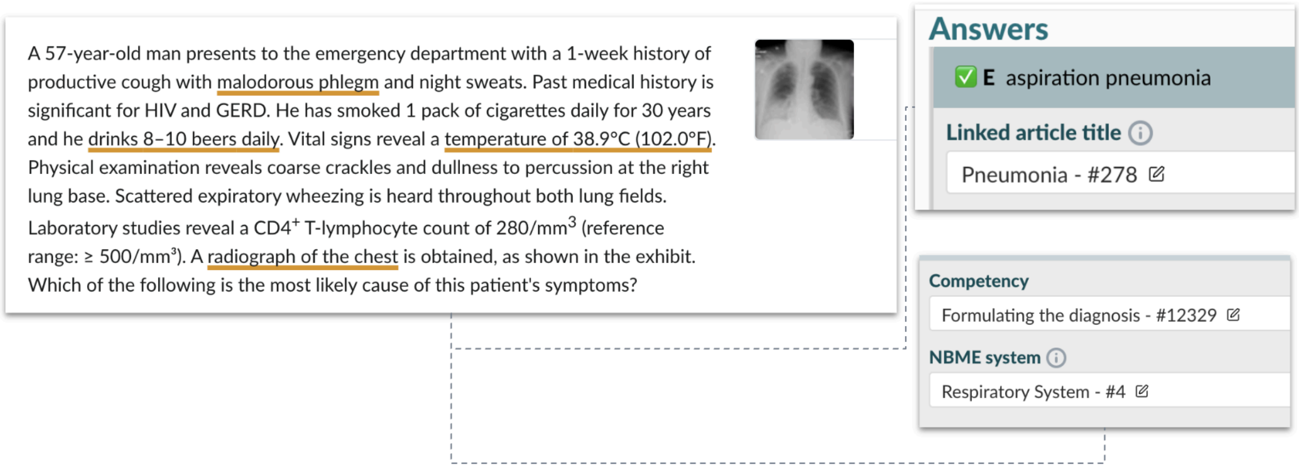

Figure 2. ***AMBOSS question stem tagged with skills.***

Leveraging prior response history and these rich skill tags, the AMBOSS platform uses the Rasch model to independently estimate a learner's knowledge on fine-grained medical topics as well as cross-cutting competencies at a given point in time. As the platform estimates one theta per skill, the system keeps track of hundreds of theta values for each learner.

When estimating the probability of success on a new question, the model combines competency- and topic-specific evidence.

Updating knowledge over time

The learner model becomes a knowledge tracing system by updating these ability estimates as new responses are observed.

The Rasch model provides an estimate of learner ability at a given point in time, which serves as the system's prior belief. Following a new response, this estimate is updated based on the relationship between the observed outcome and the probability predicted by the learner's prior ability level and the question's difficulty level.

Learner ability estimates increase after correct responses and decrease after incorrect responses, with larger adjustments occurring when outcomes are less expected by the model. Updates are further moderated by uncertainty, such that early responses produce larger shifts in ability estimates, while later responses lead to smaller and more stable refinements as evidence accumulates over time.

Applications

The AMBOSS learner model is the foundation for both personalized question sessions for individual learners and advanced insights into their exam preparedness.

Adaptive Qbank sessions

When a student starts an Adaptive Qbank session, the platform retrieves prior response history and updates the student’s ability estimates (theta values) across skills using the learner model. These thetas are used to identify the learner's weakest topics and competencies, as well as predict the probability that the learner will correctly answer unseen questions.

The selection of the first question in an Adaptive Qbank session illustrates the underlying personalization logic. The system sorts questions by the likelihood that a learner can answer questions correctly. Lower predicted probabilities indicate weaker learner proficiency in that skill. The first question delivered in an Adaptive Qbank session targets the learner’s weakest area.

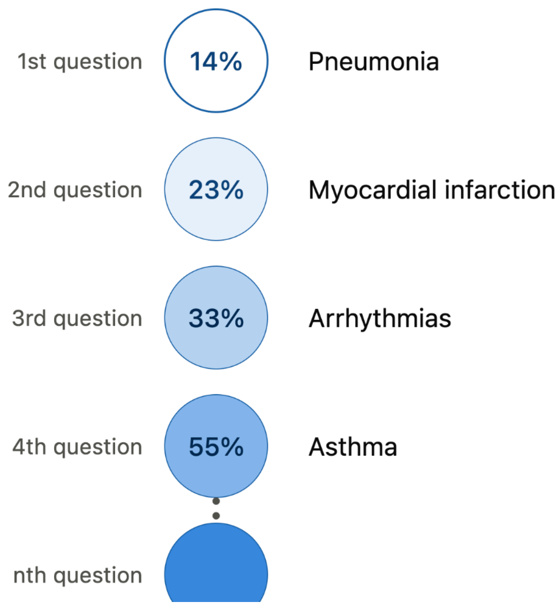

Figure 3. ***Adaptive Qbank targets learner weaknesses first.*** Each circle represents a question tagged with a skill and the probability a learner can answer it correctly.

For subsequent questions in an Adaptive Qbank session, the same process is repeated. Before selecting the next question, the learner model simulates a correct response to the previously selected question, updates the corresponding thetas, and recalculates the predicted probabilities for the remaining unseen questions. This look-ahead strategy identifies the question that is expected to provide the greatest educational value if the learner answers the prior question correctly.

This iterative process produces a personalized sequence of questions that efficiently targets the learner's knowledge gaps.

Advanced Insights

So far, we have examined how the knowledge tracing system is leveraged for question sequencing. The same system also produces advanced metrics that accurately predict student exam outcomes.

As learners complete question bank sessions, the system continuously updates their estimated proficiency (thetas) across medical topics and clinical competencies. These thetas are derived solely from completed questions and form the learner model. This learner model is then projected across every question relevant to the learner's study objective (e.g., USMLE Step 2 CK), including questions the learner has not yet attempted. For each eligible question, the system estimates the probability that the learner would answer it correctly if encountered today.

Aggregating these predicted probabilities across the entire study objective question pool produces a real-time representation of the learner’s expected performance. From this representation, the platform derives several learner-facing metrics.

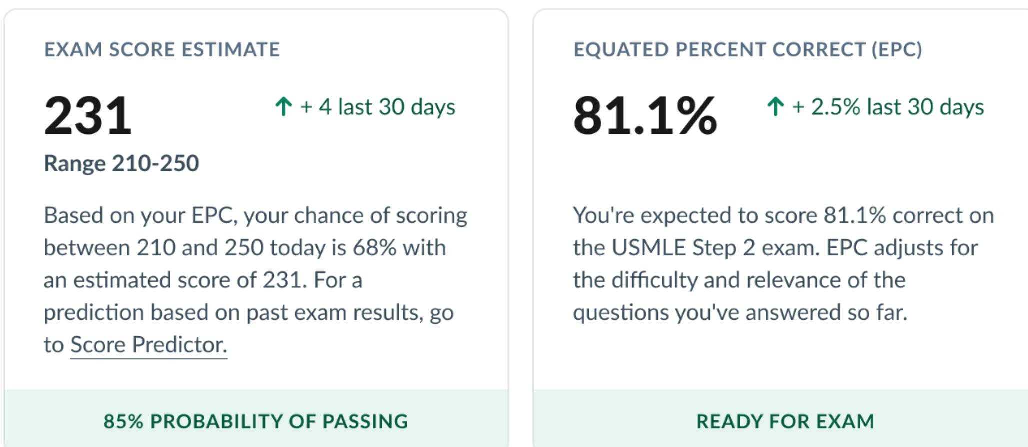

Figure 4. Metrics derived from learner ability estimates.

One such metric is the Equated Percent Correct (EPC), which represents the learner’s average predicted success rate across all questions relevant to their study objective. Unlike raw percent correct, which only reflects previously attempted questions, EPC provides a more stable estimate of learner knowledge while accounting for differences in question difficulty. In addition, as exam-relevant topics are represented by more items in the AMBOSS question bank, they naturally have greater influence on the EPC.

Building on the EPC, the platform also generates an Exam Score Estimate, representing the score a learner would be expected to achieve if they took the exam today. This estimate is derived by mapping the EPC onto the exam's official scoring scale using a model calibrated on historical data linking predicted performance to observed exam scores. The Exam Score Estimate is accompanied by a confidence interval that reflects the inherent uncertainty in predicting a learner's future exam performance

Together, these metrics translate the learner model into interpretable measures of exam readiness. They enable learners to track meaningful changes in performance over time, including trends observed during the preceding 12 months.

Validation

As both Adaptive Qbank sessions and advanced insights depend on the accuracy of the underlying learner model, we evaluated how well metrics derived from the model's estimates correspond to real-world exam performance.

For this analysis, we compared the AMBOSS exam score estimates against self-reported USMLE Step 2 CK scores from 7,244 learners. The AMBOSS exam score estimates achieved a mean absolute error (MAE) of 8.52 score points. In other words, the average amount by which AMBOSS under- or overpredicts the Step 2 CK scores is only 8.52 points.

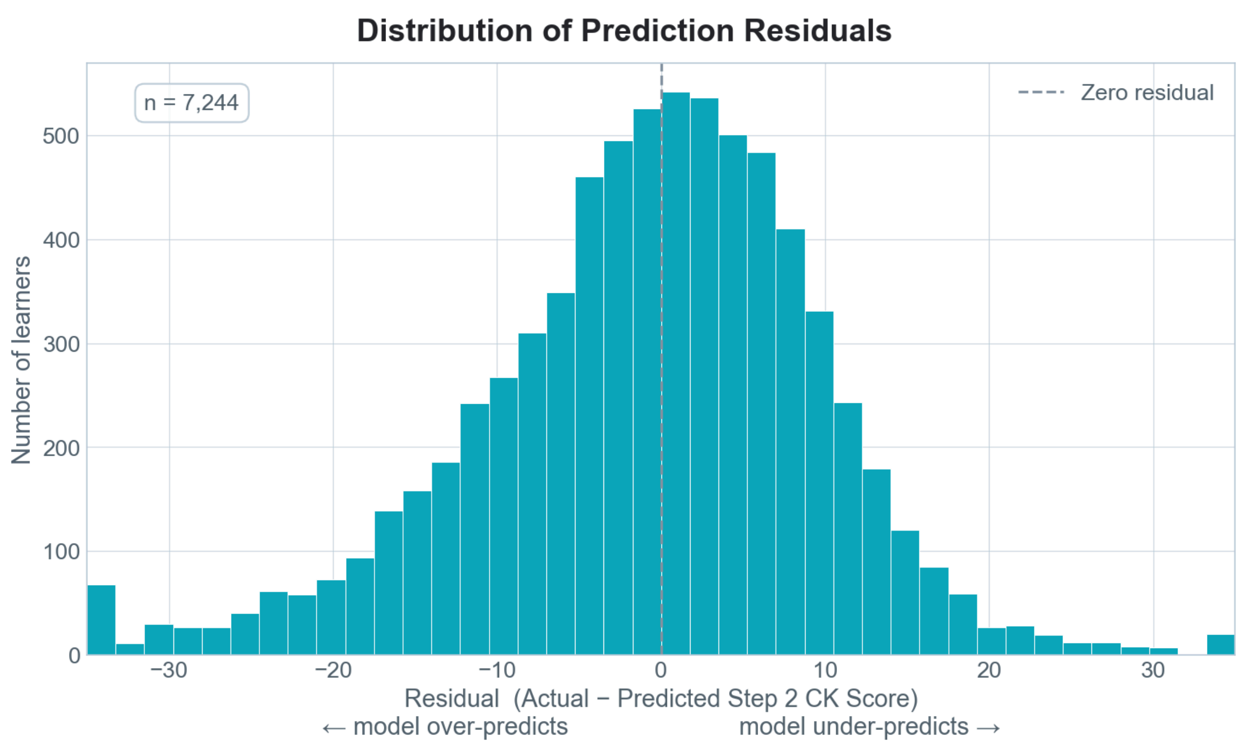

Figure 5. **How close AMBOSS predictions come to real exam scores**. Each bar counts how many of the 7,244 learners had a prediction that missed their actual Step 2 CK score by a given amount. Bars to the left are learners for whom AMBOSS predicted too high; bars to the right, too low; the dashed line in the middle marks a perfect prediction. The error concentration in the center shows that for most learners the estimate came within a few points of their real score. The roughly symmetrical bell curve shape shows AMBOSS is about equally likely to guess a little high as a little low — there is no systematic bias in estimates.

For comparison, we evaluated a naive baseline that ignores all learner-specific information and predicts the same score for every learner as the average USMLE Step 2 CK score observed in the dataset. This baseline achieved a MAE of 11.01 points. By incorporating learner-specific knowledge estimates derived from question bank activity, the AMBOSS learner model reduced prediction error by approximately 23%.

Predictive accuracy is naturally limited by a test's inherent variability. This lower bound arises from the standard error of the estimate (SEE), which represents the expected variation in a student's score if they were to take multiple versions of the exam without any changes in their knowledge or preparation.

For the Step 2 CK exam, the updated USMLE Score Interpretation Guidelines report a SEE of 8 points.[7] Based on this SEE, we calculate an MAE of 6.4 points represents the theoretical limit of predictive accuracy, as it accounts for unavoidable variation in student exam performance (see Appendix for details).

The AMBOSS exam score estimates' MAE of 8.52 is within approximately two points of this theoretical floor, demonstrating that the learner model is approaching the practical limit of prediction accuracy. This result is particularly notable because the estimates are derived from routine study data (i.e. question bank activity) rather than dedicated mock exams.

Outlook

The current learner model estimates knowledge from longitudinal question bank performance, but this represents only the beginning of a broader learning intelligence platform.

Because the model is built around underlying medical topics and competencies, not a specific exam, it provides a persistent representation of learner knowledge that can continue to accumulate evidence throughout a learner's educational journey. As learners transition between study objectives, for example from Step 1 to shelf exams or from Step 3 to specialty board exams, the current version of the learner model already carries forward accumulated evidence about what they know instead of starting from zero.

Future iterations of the learner model will incorporate additional sources of learning evidence beyond question responses. Interactions with AI experiences, free-text explanations, and other learning activities all provide signals about what a learner knows and where misconceptions remain. Integrating these diverse observations into a unified learner model will enable increasingly granular estimates of proficiency.

For educators and institutions, the learner model can create the foundation for **programmatic assessment** at scale.[8] Instead of relying primarily on isolated high-stakes examinations, institutions could aggregate thousands of low-stakes observations into a continuously evolving picture of learner development. AMBOSS plans to enable mapping learner knowledge directly onto the curriculum, allowing educators to identify struggling learners earlier, detect concepts that entire cohorts find challenging, evaluate curriculum effectiveness, and deliver targeted remediation.

Finally, AMBOSS will leverage the learner model to enable a new generation of personalized AI learning experiences, allowing students to receive coaching and study plans grounded in their unique, evidence-based proficiency profiles rather than static, one-size-fits-all learning paths.

Appendix

This section details the mathematical formulas for the Rasch model and the theoretical accuracy limits for USMLE Step 2 CK score predictions.

The Rasch model is given by a one-parameter logistic formula:

$$P(X_{ij} = 1 \mid \theta_i,\ b_j) = \frac{1}{1 + e^{-(\theta_i - b_j)}}$$

where:

- the probability of a correct response (x = 1) is a non-linear function of learner *i*'s ability level theta (θ) and item *j*'s difficulty level beta (b).
- Here, an item is a question from the AMBOSS question bank.

Mathematically, the lowest possible mean absolute error (MAE) for any prediction model for the Step 2 UK exam can be approximated as:

$$MAE_{min} \approx SEE \times \sqrt{\frac{2}{\pi}}$$

where:

- $SEE = 8$, as reported in the updated USMLE Score Interpretation Guidelines,[7]
- The factor $\sqrt{\frac{2}{\pi}} \approx 0.7979$ accounts for the expected absolute deviation from the mean in a normal distribution.

Substituting the given SEE value:

$$MAE_{min} \approx 8 \times 0.7979 \approx 6.4$$

Any model achieving an MAE close to this threshold can be considered highly accurate, as further improvements would be constrained by the inherent randomness in test-taking outcomes.

References

1. Ogundiya O, Rahman TJ, Valnarov-Boulter I, Young TM. *Looking Back on Digital Medical Education Over the Last 25 Years and Looking to the Future: Narrative Review*. J Med Internet Res. (2024). doi:10.2196/60312
2. Triola, M., and J. Burk-Rafel. *Precision Medical Education*. Academic Medicine. (2023) https://doi.org/10.1097/ACM.0000000000005227
3. Corbett, A.T., Anderson, J.R. *Knowledge tracing: Modeling the acquisition of procedural knowledge*. User Model User-Adap Inter 4, 253–278 (1994). https://doi.org/10.1007/BF01099821
4. Ghodai Abdelrahman, Qing Wang, and Bernardo Nunes. 2023. *Knowledge Tracing: A Survey*. ACM Comput. Surv. 55, 11, Article 224 (2023). https://doi.org/10.1145/3569576
5. Lord, F.M. *The Relation of Test Score to the Trait Underlying the Test*. Educational and Psychological Measurement, 13, 517 - 549 (1952). https://doi.org/10.1177/001316445301300401
6. Rasch, G. . *Probabilistic models for some intelligence and attainment tests* (Expanded ed.). University of Chicago Press (reprinted 1980).
7. USMLE Score Interpretation Guidelines. United States Medical Licensing Examination. (2022, updated 2026). https://www.usmle.org/sites/default/files/2022-05/USMLE%20Step%20Examination%20Score%20Interpretation%20Guidelines_5_24_22_0.pdf
8. Schuwirth LWT, Van der Vleuten CPM. *Programmatic assessment: From assessment of learning to assessment for learning*. Med Teach. 33(6), 478–485 (2011). https://doi.org/10.3109/0142159X.2011.565828