Journal of Education and Development; Vol. 2, No. 1; April, 2018
ISSN 2529-7996 E-ISSN 2591-7250
Published by July Press

# Integration of Learning Outcomes and CBT Data Facilitates Individual Medical Student Support in Basic Medical Education

Na Jin Kim[1] & Su Young Kim[1,2]

[1] MASTER Center for Medical Education Support, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

[2] Department of Pathology, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea

Correspondence: Su Young Kim, Department of Pathology, College of Medicine, The Catholic University of Korea, Seoul, Republic of Korea. Tel: 82-2-2258-7315.

Received: March 30, 2018 Accepted: April 18, 2018 Online Published: April 20, 2018

doi:10.20849/jed.v2i1.367 URL: https://doi.org/10.20849/jed.v2i1.367

**Abstract**

The purpose of this study was to manage each course in the spectrum of curriculum and to monitor the students' individual sequential performances. We developed a relational database containing data on learning outcomes, computer-based test results, tutors, and students. In addition, we restructured the students' test results in terms of learning outcomes to produce competency portfolios of individual students. We found that test reliabilities were variable among the courses. Reuse of test items significantly increased the difficulty and compromised the discrimination of items. By building the relationships among the items, the test results and learning outcomes helped us to monitor and manage each test. And it provided not only the students' status but their sequential changes. This made it possible to clearly visualize each individual student's current and past competencies. By building the relational database, we could evaluate diverse aspects of our curriculum and we got an idea to improve it. The database will help us to guide data-driven curriculum management to complement pedagogy-driven medical education. This would be an effective tool to manage undergraduate medical education in an integrated manner.

**Keywords:** undergraduate medical education, computer-based test, academic database, learning outcome, individual learning, student assessment

## 1. Introduction

Undergraduate medical education includes a huge network of courses on basic medical sciences, clinical sciences, and medical humanities. Traditionally, courses in undergraduate medical education were discipline-based. For example, students learned pathology in a pathology course. Since the introduction of integrated education, courses in undergraduate medical education became diverse and complex, in terms of contents and participants (Brauer & Ferguson, 2015).

Integrated education facilitated learning in the student's aspect. On the contrary, it made it more difficult to construct and manage the courses in the context of curriculum on the administration side. The Catholic University of Korea, College of Medicine changed the undergraduate medical education curriculum in 2009. Before 2009, each medical discipline was taught in a separate course. After medical education curriculum reform, we adopted integrated multidisciplinary courses throughout the curriculum. In addition, computer-based test (CBT) was adopted as the main student assessment tool in more than two-thirds of the courses.

Computers are used in education for many purposes ranging from presenting the course contents to student assessment. CBT has become one of the most common forms of testing (Akdemir & Oguz, 2008). CBTs have several strong points compared to paper and pencil-based tests (PBTs). The benefits of CBTs are well documented in several articles (Cantillon, Irish, & Sales, 2004): prompt performance feedbacks by automated marking systems, statistical analysis of the data, diverse formats of test items, and so on. Besides, CBTs are known to provide equal test scores comparable to PBTs (Hochlehnert et al., 2011). By maximizing the benefits of such CBT, CBT should be implemented in the direction of improving students' learning ability (Smoline, 2008).

In CBTs in our college, the main type of question is the multiple-choice question (MCQ). Because MCQs greatly reduce the time and effort required to manage tests and provide prompt feedback to students, they are commonly used for assessing a large group of students. However, MCQs are difficult to write, can be cueing, and may lead to creation of false knowledge (Epstein, 2007; Roediger & Marsh, 2005). Although we have accumulated tremendous amount of data on the CBT server, we did not have an opportunity to use the data other than individual score feedback in each course. One of the purposes of this work is to gain useful teaching and learning information which is not feasible to obtain from individual test results.

Generally, feedback on a single test is delivered to students at the end of every test. It is not easy to comprehend cumulative competencies of each individual student in terms of learning outcomes. Because the assessment results keep getting accumulated in the servers, the data may be a valuable source to build competency portfolios of individual students. Competency portfolio is a learning management tool (Jane, 2009). If we can analyze competency portfolios in terms of learning outcomes or some other significant keywords, it would be possible to monitor competency buildup of individual students. Besides, the portfolio makes it possible to identify and help the students who are failing in the early stage.

Our college runs about 100 courses of the undergraduate medical education curriculum. Because all the courses are connected, and many courses are multidisciplinary, it is a formidable task to maintain a balance among the courses and to monitor the quality of the courses. To better understand what we are doing in education practice, we designed a relational database that covered most of the data produced in education practice, including learning outcomes, class schedules, instructors, students, surveys, assessments and their results. Besides the data from our own college, the database included the nation-wide standards, including the learning outcomes presented by the Korean Association of Medical Colleges (KAMC) and the assessment items mentioned by the National Health Personnel Licensing Examination Board (NHPLEB) for comparison.

This article provides an example of developing all-in-one database and learning analytics extracted from the database in the setting of undergraduate medical education. Although there are many articles on specific subjects, such as learning outcomes, student portfolios, and performance assessment, research analyzing them in a holistic manner is limited. We presented a database that contains all possible data produced during undergraduate medical education and showed the way to produce learning analytics which is valuable for course management, competency assessment, and individual student support.

A model which used learning outcomes as a tool to assess progression was introduced previously (Harden, 2007). According to this model, progression of an individual student can be assessed in terms of learning outcomes. However, a systematic approach to develop such model has not been introduced to date. The tools that we introduced here might be the first curriculum-wide tools for implementing the model.

## 2. Method

### *2.1 Data Collection*

Since 2009, we have been using 3 different systems in undergraduate medical education. These include the student information system, the learning management system, and the CBT system. Students access the systems through a browser and the results are collected and deposited in the systems. For the research purpose, selected data was migrated to the research database and the data was merged with the previously reported learning outcome data (Kim et al., 2015). The database covered data on students, instructors, classes, courses, learning outcomes, survey results, and CBT. More than 25,000 test items used in 437 exams were included. To categorize the test items in multiple ways, metadata for each item was manually created. The metadata included relationships with instructors, students, learning outcomes, diseases, disciplines, course, and keywords.

The protocol for data collection and analysis was approved by the institutional review board of the Catholic University of Korea, College of Medicine (MC15EISI0121).

### *2.2 Database Construction*

We used FileMaker Pro 15 (FileMaker Inc., CA, USA) for database construction (Kim et al., 2015). We reviewed the curriculum inventory standard introduced by others and included it in the schema of the database (Ellaway et al., 2014). We included events (lectures and tests) and expectations (learning outcomes and competencies) and defined their relationships.

### *2.3 Analysis Tools*

We used the internal statistical functions of FileMaker Pro 15, Matlab 2016b (MathWorks, MA, USA) and Excel 2016 (Microsoft, WA, USA) for statistics and graph production.

## 3. Results

### *3.1 Database Construction*

Currently, the database harbors 53 tables, 299 fields, and 303,541 records. The interconnected tables can be divided into the following 6 domains: courses, surveys, students, tests, instructors and classes, and learning outcomes. The relationships among the tables created a large network of data. With the help of the information on the relationship, any data in a certain table can be tracked by others in different tables (Figure 1).

### *3.2 Evaluation of Examinations*

Using the data from 437 CBTs, we calculated the reliability of each test, average .67 (SD = .15). Most of the reliability coefficients were .6 or higher. However, 4 courses showed low levels of reliability coefficients (< .5). To identify the possible causes of low scores, we calculated the correlation coefficients between reliability coefficients and the number of test items in each test (Figure 2). The reliability coefficients were significantly correlated with the number of test items in each test ($r = .64$, $P < .001$). Although it is already known that the size of the test items is related to its reliability, we found that at least 38 test items were required to maintain the reliability at the level of .6 in our school.

### *3.3 Characteristics of the Test Items Used*

To determine whether we are assessing the students in the appropriate way, we categorized the test items according to the type and content of the questions and calculated the number of items in each category. We found that the test items placed a disproportionate emphasis on diagnosis, memory recall, and certain clinical presentations (CPs). Among the 105 CPs that were included in the curriculum, 9 CPs were not used for student assessment.

To evaluate the quality of test items, we calculated the difficulty and discrimination of each item. The means of difficulty and discrimination index were .73 (SD = .26) and .26 (SD = .21), respectively. A type (one best-answer) items showed higher difficulty index; .74 (SD = .26), than R type (extended-matching) and short-answer type questions. R type items showed higher discrimination index; .3 (SD = .16), than other items. We did not find a significant difference in difficulty and discrimination among item types related to memory recall, data interpretation, and problem solving.

After a manual review of test items, we found that some of these items were used in the previous years and other courses. Although we guide tutors to develop new test items, some of the items are similar to or same as the ones used previously. Because students might have a chance to share the information on assessments, we wondered if the reused item had any effect on the assessment.

We divided the reused items into 2 groups. The same item group was defined by identical question stems and identical multiple options. The similar item group was defined by the same clinical vignette in a question stems and different multiple options.

We reviewed 25,131 MCQs used between 2009 and 2014. We found 5,017 same test items and 1,424 similar test items. The difficulty increased when the test items were reused, and the tendency was greater in the same item group (Figures 3a and 3b). On the contrary, discernible changes in discrimination were not found (Figures 3c and 3d).

To determine if the changes in difficulty and discrimination of test items were statistically significant, we tested the hypothesis that the difficulty and discrimination were the same after item reuse with use of Wilcoxon signed rank test (Table 1). We found that reuse of the same items significantly increased the difficulty and decreased the discrimination of the items. Although reuse of a similar item significantly increased the difficulty, the changes in discrimination were not significant. Therefore, reuse of the same item should be discouraged to maintain good discrimination power of student assessments.

### *3.4 Competency Portfolios for Individual Students*

By building the relationships among the items, the test results and learning outcomes helped us to monitor and manage each test. Traditionally, the students receive their test results after each test and at the end of the course. It is not easy to track individual students' buildup of competencies per learning outcomes. To better understand and provide the students' status and sequential changes, we restructured the test results based on learning outcomes, CP, and keywords. This made it possible to clearly visualize the current and past competencies of each student (Figure 4). This information would be more helpful for instructors and students to focus on their weak competencies in an environment with limited resources.

Table 1. Statistics of difficulty and discrimination changes

| | Difficulty (mean) | | | Discrimination (mean) | | |
|---|---|---|---|---|---|---|
| | Initial | Later | P value* | Initial | Later | P value* |
| The same item (n = 3,209) | .75 | .90 | < .001** | .28 | .25 | < .001** |
| A similar item (n = 1,082) | .73 | .78 | < .001** | .27 | .28 | .65 |

* Comparison of the initial and the latter results by Wilcoxon signed rank test.

** Statistically significant.

## 4. Discussion

In this article, we showed how we integrated the data on undergraduate medical education. Using the integrated data, we could evaluate diverse aspects of student assessment and provide individualized feedback on individual competencies to medical students.

For proper student assessment, comparison of test items with the ones from other courses or previous school years is highly recommended. We showed that item reuse significantly compromised the indices of difficulty and discrimination. There are several articles on item reuse with variable results (Mills, 2002; Wollack & Fremer, 2013; Wood, 2009). Because, our results are obtained from a large cohort, over the relatively long period, and using a curriculum-wide approach, we believe that the results are convincing.

Frequent feedback is important in student success (Tinto, 2012). If the feedback is timely and focused on competencies of individual students, the effect would be maximized. The approach introduced in this article facilitated frequent and prompt feedback on individual competencies to medical students.

Assessment drives learning (Wass et al., 2001). Because the students pay attention to the contents of the tests, carefully designed assessments themselves facilitate student learning even without feedback (Epstein, 2007). The driving force of each test decreases after the test. In addition, it is difficult for students to build their cumulative results of the tests that they have taken in many courses. With a competency portfolio that we suggested in this work, students may derive benefit and further it may guide their learning in a balanced manner.

Analysis of learning data may invoke ethical issues. Data on student activities and performances contains personal information. Feedback based on individual information should be restricted to limited stakeholders. Any attempts to obtain information on gross perspectives should be based on anonymized data separated from the working data containing students' personal information.

Presenting undergraduate medical education in a digital format is a kind of difficult task. Significant portion of education practice may not be transferable to a database. We need to decide which data is to be included and to understand the limit of the measured values in this process. Therefore, learning analytics based on DBs does not reflect all aspects of education practice. Some of the courses do not rely on CBT and LMS for teaching and learning, and it is difficult or impossible to measure direct personal communication in a digitally recordable format. Therefore, the data stored in DBs is inevitably limited.

Although we made every effort to accommodate all possible data into the database, significant portion of medical education is missing in the database. Miller's framework for clinical assessment consists of "knows", "knows how", "shows how", and "does" (Miller, 1990). Because CBTs assessed the first 2 levels, it is not possible to evaluate students' competency on "shows how" and "does". Most of the assessments of clinical skills are based on a standard patient interview, simulation, and face-to-face evaluations. Therefore, the assessment results of clinical skills are not covered in the current work. However, the competency portfolio produced by our approach may be an informative measure for the students at the beginning of clinical clerkship. We will continuously work to develop more effective ways to integrate clinical skill assessments with other digitally recorded data to complete our work.

Although there are globally accepted standards in undergraduate medical education, each medical school has its own mission that differs from the missions of the other medical schools. General management systems may not be adequate to follow up their own goals. Therefore, every school needs to develop school-specific solutions and analytics in their own education environment. Well-designed and properly-managed education systems will provide medical schools the flexibility to accommodate ever-changing international standards and individual support for medical students with variable performance.

With the recent advancement in medical sciences, the demand for diverse contents has increased in medical education. The number of students is increasing. Using the tools and the approach we introduced here, it would be possible to manage the complex medical education curriculum. At the same time, we can support individual students efficiently.

By building a relational database, we could evaluate diverse aspects of our curriculum and we got an idea to improve it. The database will help us to guide data-driven curriculum management to complement pedagogy-driven medical education. We hope the experience presented in this article may inspire other medical schools to start learning analytics for quality improvement and to provide personalized feedbacks to medical students.

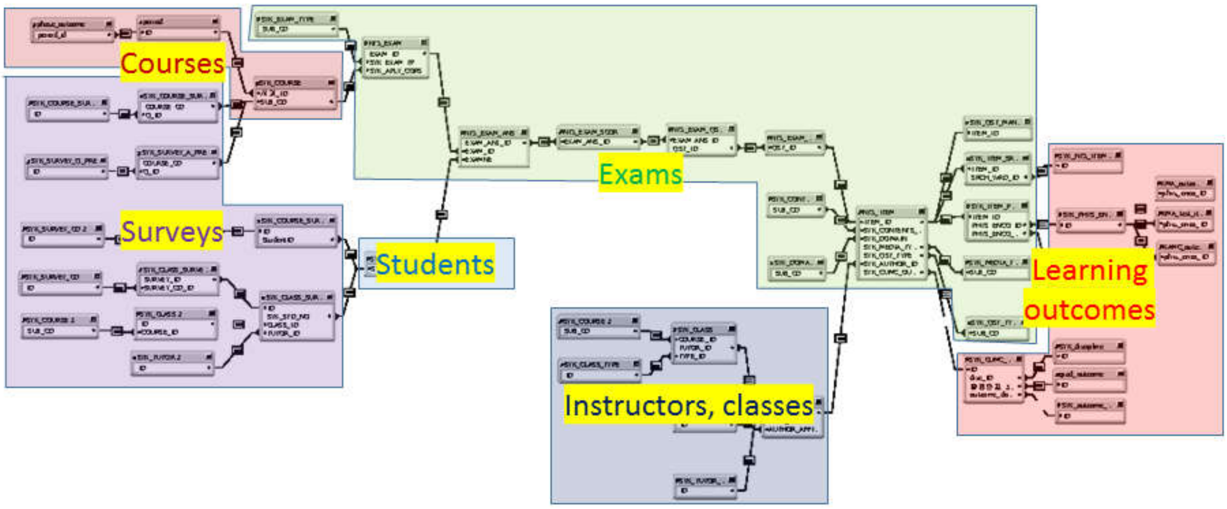

Figure 1. Relationships of tables containing medical education data in a database

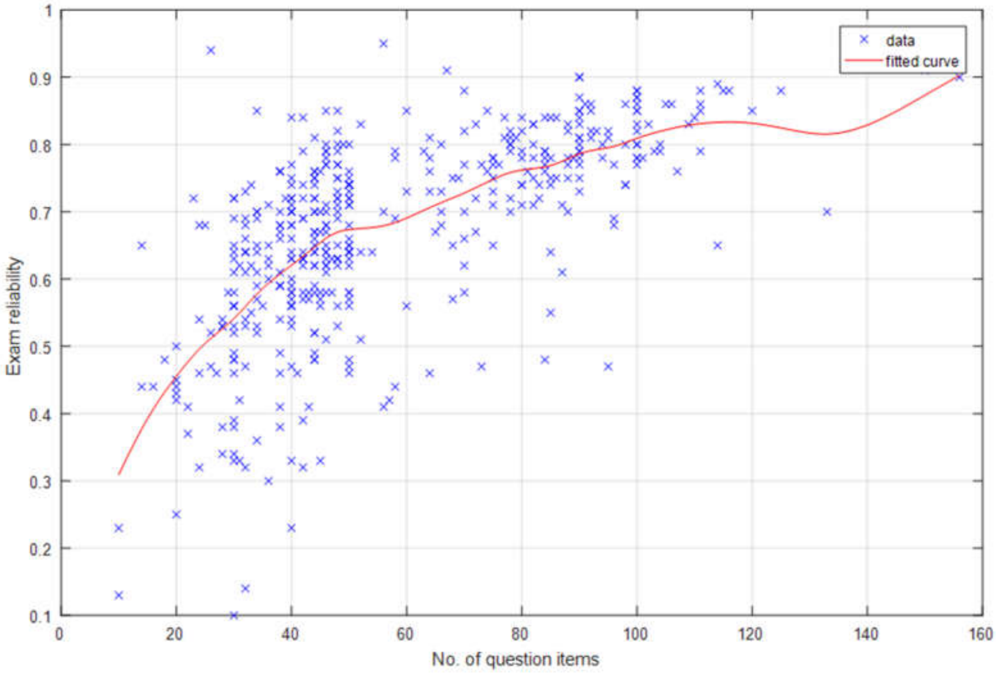

Figure 2. Scatter plot showing test reliability versus number of test items (n = 437 tests) (the correlation coefficient is .64 (P < .001).)

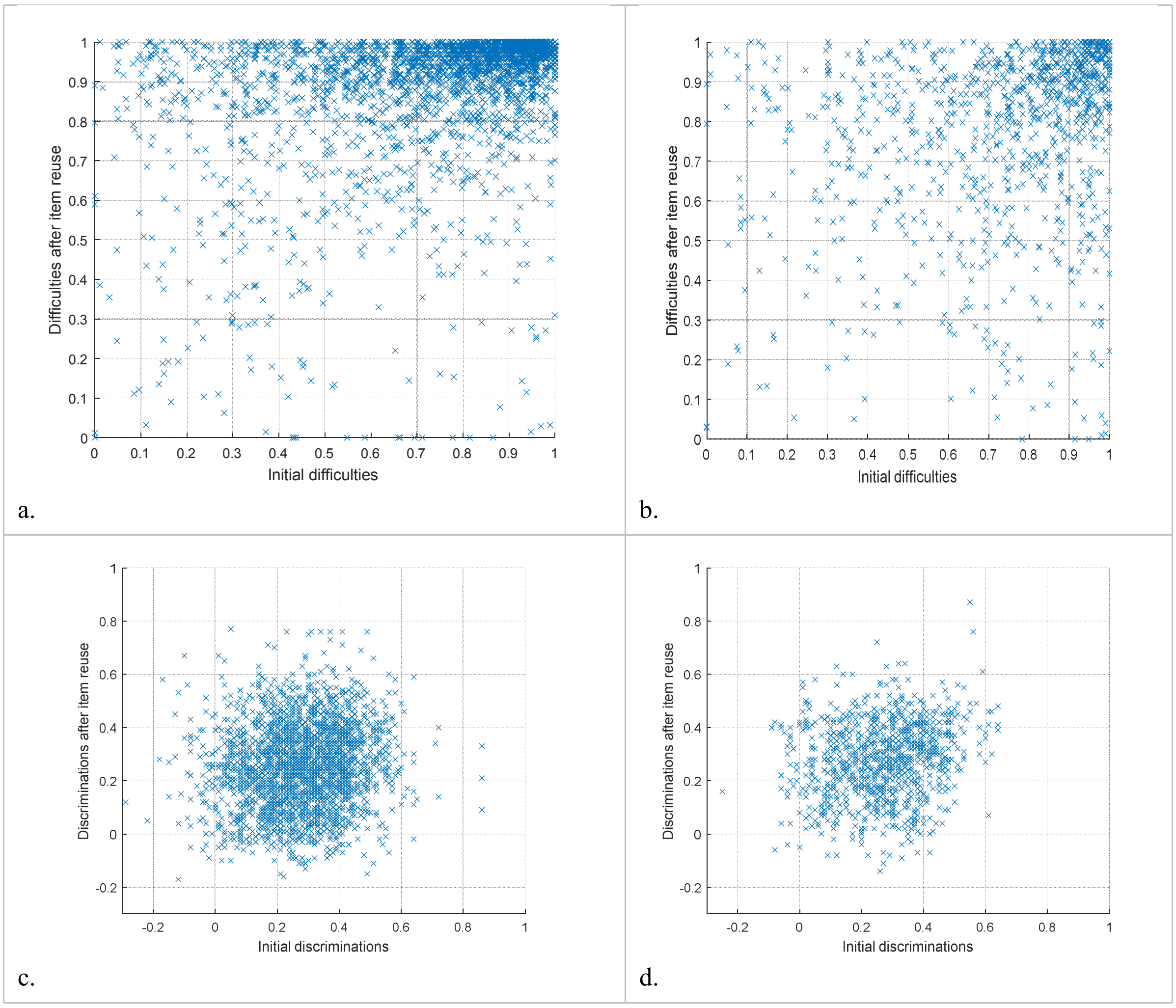

Figure 3. Changes in difficulty (a and b) and discrimination (c and d) caused by reuse of the same test item (a and c) and a similar (b and d) test item

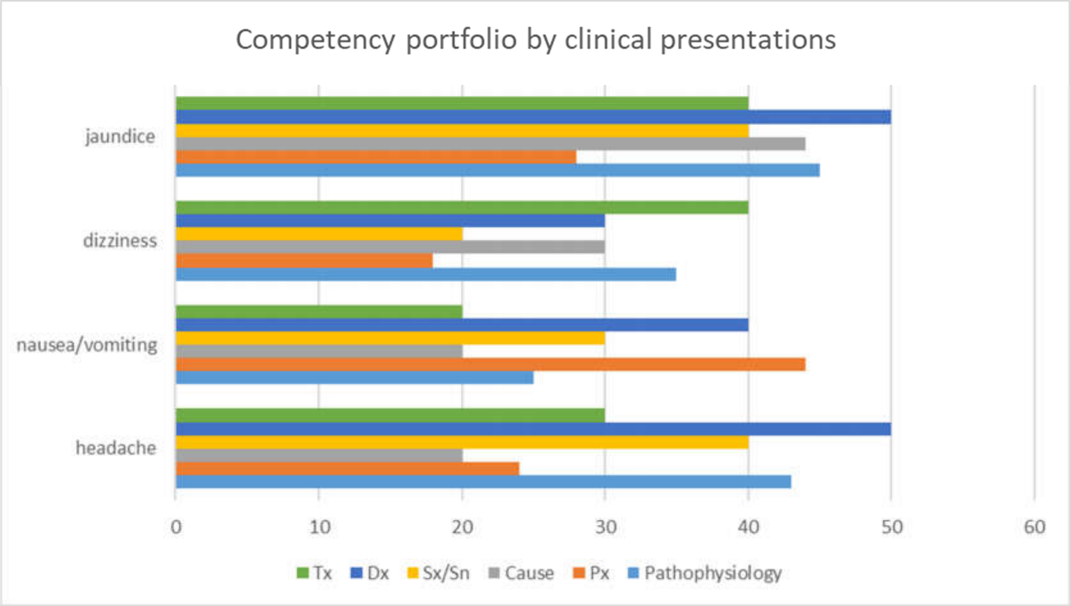

Figure 4. An example of a competency portfolio for measuring student competency (all assessment results of a student were restructured per clinical presentation.)

## References

Akdemir, O., & Oguz, A. (2008). Computer-based testing: An alternative for the assessment of Turkish undergraduate students. *Computers & Education, 51*(3), 1198-1204. https://doi.org/10.1016/j.compedu.2007.11.007

Brauer, D. G., & K. J. Ferguson. (2015). The integrated curriculum in medical education: AMEE Guide No. 96. *Med Teach, 37*(4), 312-22. https://doi.org/10.3109/0142159X.2014.970998

Cantillon, P., B. Irish, & D. Sales. (2004). Using computers for assessment in medicine. *British Medical Journal, 329*(7466), 606-609. https://doi.org/10.1136/bmj.329.7466.606

Ellaway, R. H., S. Albright, V. Smothers, T. Cameron, & T. Willett. (2014). Curriculum inventory: Modeling, sharing and comparing medical education programs. *Med Teach, 36*(3), 208-15. https://doi.org/10.3109/0142159X.2014.874552

Epstein, R. M. (2007). Medical education - Assessment in medical education. *New England Journal of Medicine, 356*(4), 387-396. https://doi.org/10.1056/NEJMra054784

Harden, R. M. (2007). Learning outcomes as a tool to assess progression. *Med Teach, 29*(7), 678-82. https://doi.org/10.1080/01421590701729955

Hochlehnert, A., K. Brass, A. Moeltner, & J. Juenger. (2011). Does Medical Students' Preference of Test Format (Computer-based vs. Paper-based) have an Influence on Performance? *Bmc Medical Education, 11*. https://doi.org/10.1186/1472-6920-11-89

Jane, M. (2009). *The Competency Portfolio as a Learning Management Tool*. Toronto: CAPLA Fall Focus Workshop.

Kim, Na Jin, In Ae Park, Eum Ju Kim, Seung Ae Baek, Nani Kwon, Hye In Lee, & Su Young Kim. (2015). Evaluation of Concordance between Learning Outcomes of Basic Medical Education Courses and Assessment Items of the Medical Licensing Examination. *Korean Medical Education Review, 17*(1), 33-38. https://doi.org/10.17496/kmer.2015.17.1.33

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Acad Med, 65*(9 Suppl), S63-7. https://doi.org/10.1097/00001888-199009000-00045

Mills, Craig N. (2002). *Computer-based testing: building the foundation for future assessments*. Mahwah, N.J.: L. Erlbaum Associates.

Roediger, H. L., 3rd, & E. J. Marsh. (2005). The positive and negative consequences of multiple-choice testing. *J Exp Psychol Learn Mem Cogn, 31*(5), 1155-9. https://doi.org/10.1037/0278-7393.31.5.1155

Smoline, D. V. (2008). Some problems of computer-aided testing and “interview-liketests”. *Computers & Education, 51*(2), 745-756. https://doi.org/10.1016/j.compedu.2007.07.008

Tinto, Vincent. (2012). *Completing college: rethinking institutional action*. Chicago, London: The University of Chicago Press. https://doi.org/10.7208/chicago/9780226804545.001.0001

Wass, V., C. Van der Vleuten, J. Shatzer, & R. Jones. (2001). Assessment of clinical competence. *Lancet, 357*(9260), 945-949. https://doi.org/10.1016/S0140-6736(00)04221-5

Wollack, James A., & John Fremer. (2013). *Handbook of test security*. New York: Routledge.

Wood, T. J. (2009). The effect of reused questions on repeat examinees. *Advances in Health Sciences Education, 14*(4), 465-473. https://doi.org/10.1007/s10459-008-9129-z

## Copyrights