Iyer, S. et al. (Eds.) (2022). Proceedings of the 30th International Conference on Computers in Education.
Asia-Pacific Society for Computers in Education

Combining Data and Human Intelligence through Predictive Visual Analytics to Improve Educational Assessments

Yancy Vance PAREDES[a*] & I-Han HSIAO[b]
[a]*Arizona State University, USA*
[b]*Santa Clara University, USA*
*yvmparedes@asu.edu

Abstract: Assessments are widely used in higher education. It is often the case that most teachers learn the tacit knowledge of creating tests throughout their careers. Despite the proliferation of educational technologies, existing datasets are often still untapped for their potential use to improve test quality. This paper proposes a novel approach to leveraging students' test data to aid teachers in the test construction process. Based on the various student profiles and the domain model learned by the system, teachers can administer their newly created tests through simulation, effectively helping them identify items for revision and improvement without additional effort. The affordances of interactive data visualization are utilized to make such predictive models accessible for the teachers. In addition to the simulation, the system can determine topic coverage, ensuring such test matches a predetermined test blueprint. This system integrates machine intelligence with human intelligence that benefits the teacher in making informed decisions. Students indirectly benefit as they are assessed with better quality tests that could capture their mastery of the domain.

Keywords: Assessments, learning analytics, knowledge tracing, human-data interaction

1. Introduction

Assessments or tests, both formative and summative, are widely used by teachers in higher education to determine their students' mastery. The outcomes of these assessments often form the basis of several important decisions affecting students. Teachers, therefore, have to ensure that these are fair, reliable, valid, and high quality. Unfortunately, there is an abundance of teachers who lack formal training on test construction. Teachers typically learn this skill from experience as they repeatedly create, administer, grade, and revise tests throughout their profession. Ideally, they become better as time progresses, but there is no guarantee. It is typically uncertain how students–even the excellent ones–would perform on a poorly devised test. Just like students, if teachers are unguided or lack feedback, they learn the process inefficiently. However, unlike students, such ineffective performance would affect those beyond themselves.

Existing techniques in psychometrics concerning the quality of a test (e.g., item response theory) often leverage student answers or responses. Several educational systems exist that can capture these data at a finer level. For example, WebPGA is a home-grown web system that streamlines the process of digitizing, grading, and distributing paper-based tests (Paredes & Hsiao, 2021). In addition to performance, it also captures behaviors relating to how students review their mistakes. These types of student data are rich and often untapped. Beyond the overall score, detailed and personalized feedback can be provided to students and teachers by building diagnostic and predictive models.

This paper proposes a novel approach to extending WebPGA by providing a tool to support teachers in test construction. This work takes inspiration from the intelligent tutoring system (ITS) literature, where students' mastery of topics is estimated based on their answers to practice opportunities (Baker, 2016). However, unlike ITS, where there is a learning component between practices, there is not generally a learning opportunity, per se, in our system as students mainly use it to access their graded tests. However, such summative tests could be framed as an opportunity to diagnose student misconceptions to tailor the subsequent tests to their capabilities.

2. Extended Research Platform

VanLehn and colleagues (1994) have outlined possible uses and benefits of simulated students in teacher education. However, most of the approaches mainly focused on instruction rather than testing. With the growing ease of collecting student test data, it is worth exploring the use of simulated students and test data in the professional development of teachers. However, this raises the question regarding the scope of the system's control. How authoritative will the system be? Will it make the decisions for the teachers? Baker (2016) articulated a paradigm shift toward designing educational systems intelligently rather than designing intelligent systems. He further argues that, ultimately, humans should make such critical decisions. Systems could help produce distilled reports that humans can use. Pelánek (2021) outlined several approaches to visualizing learner data and further notes how some of these data are often underutilized. Most of these typically pertain to student test data, such as an item's difficulty and amount of time to answer. Therefore, these prior works motivated the extension and further development of our existing system.

Existing test data is used to create the domain model (e.g., hierarchy and pre-requisite of topics) and the student profiles (e.g., multiple student models). Each student profile will have its associated proficiency parameters for each domain topic (i.e., knowledge component). The predictive model would then use these to determine the probability of a student answering the item correctly. Each item is typically composed of multiple skills. There are several ways in which student models can be developed. Item response theory (IRT) is used in psychometrics to determine students' abilities (e.g., the Rasch model). One issue with the canonical implementation of IRT is that it is unidimensional. This means students are only classified based on their overall ability, which may not be necessarily beneficial in providing a comprehensive and specific view of students' mastery. Another approach is through cognitive diagnostic models (e.g., DINO) (DiBello, Stout, & Roussos, 1995). These models are capable of associating multiple skills to a single item. However, one such limitation is that the latent features (i.e., skills) are dichotomous. This means that a given skill could either be mastered or not by the student with no varying levels of certainty. Another approach is performance factor analysis (Pavlik, Cen, & Koedinger, 2009). It is a logistic regression model that modifies the learning factors analysis. It attempts to address the issues mentioned above. In these models, student performance serves as the dependent variable. These are some of the many approaches to knowledge tracing. Liu and colleagues (2021) provided an overview of several techniques for knowledge tracing and proposed a taxonomy from a technical perspective, namely probabilistic, logistic, and deep learning. Moreover, Gervet and colleagues (2020) compared the predictive performance of these models using various educational datasets and determined the nature of the dataset on which they performed better.

Once the system learns the models, these will be stored for later instantiation in the simulation. The simulation can be configured so that tests can be administered to all the various student profiles previously seen by the system (i.e., for multiple offering courses) or those limited to the current classroom distribution only. Teachers then upload a newly created test which the system parses and automatically assigns the relevant topics. Additionally, it builds the needed inputs for the predictive models (e.g., Q-matrix). Afterward, it predicts the likelihood of simulated students answering the questions correctly (Figure 1). To further aid the process, items would be measured for similarity to make suggestions for teachers to consider. For example, a given topic is assessed multiple times, resulting in redundancy, inefficiency, or insufficient coverage.

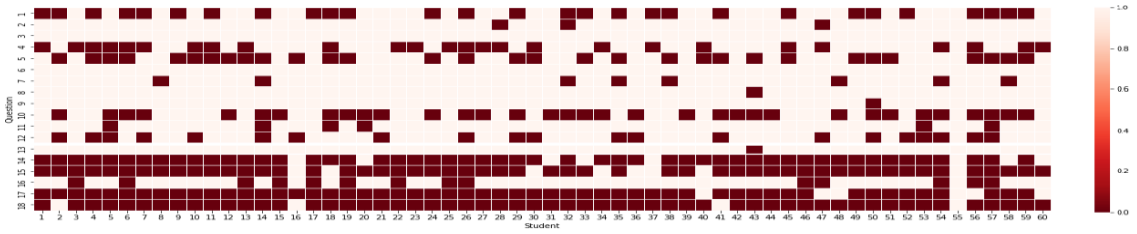

Figure 1. A prototype interactive visualization, rendered using the predictive models, illustrating the likelihood that a simulated student (x-axis) correctly answers a question (y-axis).

A typical concern of developing predictive models is their complex nature and interpretability. This could be addressed by leveraging the affordances of data visualization so that teachers are

presented with easily understood information. Following the principles of Shneiderman's (1996) Information Seeking Mantra or the "overview first, zoom and filter, then details-on-demand" (p. 337), teachers are not overwhelmed with information. Nevertheless, they are provided with a toolkit to encourage exploratory analyses. Interesting questions could be answered with the aid of the system, such as "which items are predicted to be difficult", "what topics do most students find difficult", or "what questions seem to be interdependent". The described workflow is similar to performing a modern test analysis as in psychometrics (e.g., IRT). The only difference is the timing when it is done. In this case, before a test is administered to the actual students giving the teachers a chance to improve. This knowledge could be used to objectively understand whether the students are ready to be tested or not. The created questions could be tailored to the level of performance of students. Ultimately, the teacher decides to stick with the questions or do any revisions. However, the system captures these actions for future analysis, essentially to know how this knowledge impacts teachers' decision-making.

3. Conclusion and Future Work

This work proposes a novel approach to using existing students' test data to support the process of test construction. By extending an existing educational system, insights can be provided to both students and teachers. Modeling approaches from psychometrics and artificial intelligence, such as performance factors analysis, are utilized to create various student profiles and learn the domain model. With the aid of interactive visualizations, teachers are provided with a simulation tool that allows them to forecast the outcome of a newly created test based on students' current level of mastery. Such knowledge empowers them to make informed decisions to improve the test's quality, leading to better and more effective assessments that precisely measure students' knowledge. Arguably, the role of the machine is to aid the amplification of human intelligence (Baker, 2016) with the aid of visualizations (Bassen, Howley, Fast, Mitchell, & Thille, 2018; Pelánek, 2021). In the long term, these revision behaviors of teachers could uncover best practices that can be shared with teachers early in their careers. This helps fully understand how to close the loop toward a shift to automated educational assessments.

To truly determine the system's effectiveness, it will be evaluated in two ways. First, the accuracy of the predictive models needs to be determined and compared. This could be conducted through the use of existing datasets along with the collection of new datasets. Second, the overall perceived usefulness of the system to the teachers should be determined. This will be done through subjective evaluation and surveys to solicit the teachers' feedback.

References

Baker, R. S. (2016). Stupid tutoring systems, intelligent humans. *International Journal of Artificial Intelligence in Education, 26*, 600–614.

Bassen, J., Howley, I., Fast, E., Mitchell, J., & Thille, C. (2018). OARS: Exploring instructor analytics for online learning. *Proceedings of the 5th Annual ACM Conference on Learning at Scale*.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. *Cognitively diagnostic assessment* (pp. 361–389).

Gervet, T., Koedinger, K., Schneider, J., & Mitchell, T. (2020). When is deep earning the best approach to knowledge tracing? *Journal of Educational Data Mining, 12*(3), 31–54.

Liu, Q., Shen, S., Huang, Z., Chen, E., & Zheng, Y. (2021). A survey of knowledge tracing. https://doi.org/10.48550/arXiv.2105.15106

Paredes, Y. V., & Hsiao, I.-H. (2021). WebPGA: An educational technology that supports learning by reviewing paper-based programming assessments. *Information, 12*(11), Article 450.

Pavlik, P. I., Jr, Cen, H., & Koedinger, K. R. (2009). Performance factors analysis–a new alternative to knowledge tracing. *Proceedings of the 14th International Conference on Artificial Intelligence in Education*.

Pelánek, R. (2017). Bayesian knowledge tracing, logistic models, and beyond: An overview of learner modeling techniques. *User Modeling and User-Adapted Interaction, 27*, 313–350.

Pelánek, R. (2021). Analyzing and visualizing learning data: A system designer's perspective. *Journal of Learning Analytics, 8*(2), 93–104.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings of the 1996 IEEE Symposium on Visual* Languages (pp. 336–343).

VanLehn, K., Ohlsson, S., & Nason, R. (1994). Applications of simulated students: An exploration. *Journal of Artificial Intelligence in Education, 5*(2), 135–175.