On Automated Assessments: The State of the Art and Goals

Contributions from Varun Aggarwal[*], Steven Stemler[†], Lav Varshney[‡] and Divyanshu Vats[§]
Co-organizers, ASSESS, KDD, 2014.

www.aspiringminds.com/assess/2014

This white paper is an outcome of the ASSESS workshop, which was held at KDD 2014. The paper primarily discusses why assessments are important, what is state of art and what goals should we pursue as a community. It is a brief exposition and serves as a starting point for a discussion to set the agenda for the next decade.

1. The assessment ecosystem

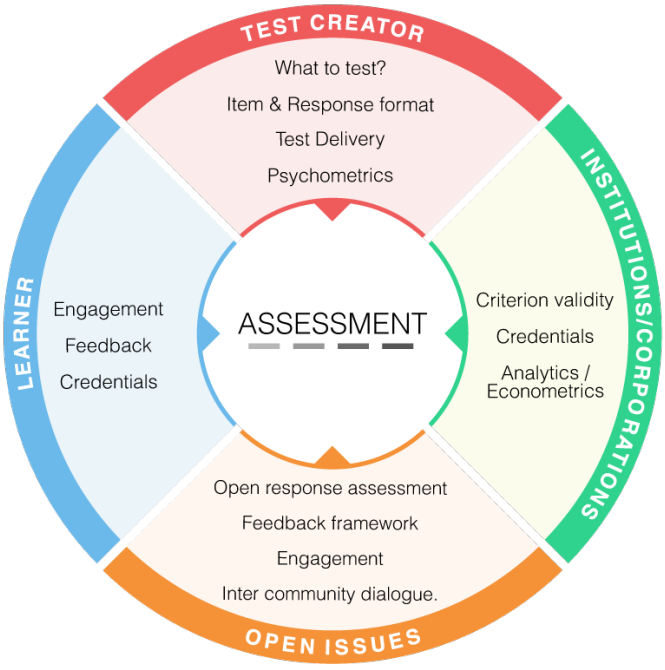

Figure 1: Learners take assessment to get feedback and credentials to show qualification. They need to be engaged. Corporations and Institutions use assessment results for recruitment decisions. We need assessments that predict, can provide standardized credentials and allow meritocracy. Test creators need to create tests keeping in mind the needs of the learners and the institutions, which ultimately leads to social good. The creators comprise education psychologist, computer scientists and subject matter experts. Some of the key open issues include automated open response assessments, engaging students to assessments, a proper framework to give feedback and the collaboration of scientific communities to make this happen.

[*]CTO, Aspiring Minds Pvt. Ltd.
[†]Associate Professor, Wesleyan University.
[‡]Assistant Professor, University of Illinois at Urbana-Champaign
[§]Postdoctoral Fellow, Rice University.

2. Why are assessments important?

Automated and semi-automated assessments are a key to scaling learning, validating pedagogical innovations, and delivering socio-economic benefits of learning.

- **Practice and Feedback**: Whether considering large-scale learning for vocational training or non-vocational education, automating delivery of high-quality content is not enough. We need to be able to automate or semi-automate assessments for formative purposes. Substantial evidence indicates that learning is enhanced through doing assignments and obtaining feedback on one's attempts. In addition, the so-called "testing effect" demonstrates that repeated testing with feedback enhances students long-term retention of information. By automating assessments, students can get real-time feedback on their learning in a way that scales with the number of students. Automated assessments may become, in some sense, "automated teaching assistants" [1].
- **Education Pedagogy**: There is a great need to understand which teaching/learning/delivery models of pedagogy are better than others, especially with new emerging modes and platforms for education [2]. To understand the impact of and compare different pedagogies, we need assessments that can summatively measure learning outcomes precisely and accurately. Without valid assessments, empirical research on learning and pedagogy becomes questionable.
- **Learning to socio-economic mobility**: For learners that seek vocational benefits, there need to be scalable ways of measuring and certifying learning so that they may garner socio-economic benefits from what they've learnt. There need to be scalable ways of measuring learning so as to predict the KSOAs (knowledge, skills and other abilities) of learners to do specific tasks [3]. This will help both learners and employers by driving meritocracy in labor markets through reduced information asymmetries and transaction costs. Matching of people to jobs can become more efficient [4].

3. What is the state of the art?

- Evolved item-response-theory based standardized and adaptive testing for high-stake assessments [5].
- Some success in automated constructed response grading and feedback generation in areas of essay grading, spoken English grading, computer program grading, and mathematical proof checking [6][7].
- Machine learning, crowdsourcing and peer-grading coupled with other computer science techniques such as program analysis show success towards scaling assessments [8][9].
- Peer assessment works when strongly driven by a rubric for assessment [5].
- Enough evidence on standardized assessments being predictive of job-success in various jobs, however predictive ability limited to 0.3-0.65 (Pearson correlation coefficient) depending on the job [10].
- Some initial work and preliminary success in automated item synthesis [11].
- Camera/microphone based remote assessment delivery solutions for mid-stake assessments [1].
- Preliminary work on gamification of assessments and with a look out evidence of validity [12].

4. What are the key problems to solve during the next five years?

- Making assessment engaging and motivating people to take it seriously [10].
- High-stake test delivery with integrity and question bank security [13].

- Constructed response problem grading for formative and high-stake purposes. To design grading techniques cannot be tricked [14][2][15].
- Assessments that could simulate the task to be tested. Create natural real-world scenarios where, for instance, test-takers not only answer questions, but also ask questions [16].
- Automated testing of personality/style/preference (without being tricked), soft-skills, behavior and practical hand-skills [17][18].
- Framework and taxonomy to provide formative feedback and find impact on learning [19][1].
- Establish how high-stake/low-stake assessment scores together with other student behavior parameters over course of learning/job are predictive of success in specific tasks and specifically employment outcomes/success [20].
- Creating games to assess, that are engaging and show high validity [21][22].
- Delivering and validating assessments through new UI and platforms such as phones, tabs and virtual reality devices.

5. What can help solve these problems?

- Greater communication between the education (I/O) psychology community, computer scientists and practitioners. Evolving a common language and finding areas of collaboration [23].
- Interest the machine learning community and others through open challenges on automated assessments of different tasks [8][9].
- Request for proposals and data challenges sponsored by corporations to engage researchers in cutting edge research in the area.
- Standard benchmark data sets to test automated grading algorithms in both trick and non-trick situations [24].
- Better collaboration and connectedness between researchers through serious social networking websites like ResearchGate and workshops.

6. Acknowledgements

Thanks a ton to Shashank Srikant and Bhanu Pratap Singh who helped to make this happen.

7. Appendix

7.1 Invited talks

- **Strapping jet engines to the stage coach: Using technology to push the boundaries of educational measurement** by ***Damian Bebell, Lynch School of Education, Boston College***
  Damian discussed the importance of continuous data collection in classrooms and feedback generated from therein.

- **Problem Generation and Feedback Generation** by ***Sumit Gulwani, Microsoft Reearch Redmond***
  Sumit presented how the problem of automatic generation of problem statements and feedback to student solutions can be cast as a problem in program synthesis. Related work in automatic generation of algebra problems, SAT-like English word problems, trigonometry problems, logic was discussed. This was followed by a discussion on feedback systems he has developed for automated computer program grading and Finite state Automata evaluation.

7.2 Papers

- **Game-Based Assessment: Two Practical Justifications** by *Thomas Heinzen*
- **Peer Mediated Testing** by *Igor Labutov, Kelvin Luu, Thorsten Joachims and Hod Lipson*
- **Boredom Across Activities, and Across the Year, within Reasoning Mind** by *William L. Miller, Ryan Baker, Matthew J. Labrum, Karen Petsche and Angela Z. Wagner*
- **Mining Student Ratings and Course Contents for Computer Science Curriculum Decisions** by *Antonio Moretti, Jose Gonzalez-Brenes, Katherine McKnight and Ansaf Salleb-Aouissi*
- **Test-Driven Synthesis for Automated Feedback for Introductory Computer Science Assignments** by *Daniel Perelman, Sumit Gulwani and Dan Grossman*
- **Some Scaling Laws for MOOC Assessments** by *Nihar Shah, Joseph Bradley, Sivaraman Balakrishnan, Abhay Parekh, Martin Wainwright and Kannan Ramchandran*

7.3 Posters

- **Classifying Peer Tutee Learning Gains with Hidden Markov Models** by *Yoav Bergner, Erin Walker and Amy Ogan*
- **Consensus Ratings: Reconceptualizing Additive Bias** by *Stephen France*
- **Communication Communities in MOOCs** by *Nabeel Gillani, Rebecca Eynon, Michael Osborne, Isis Hjorth and Stephen Roberts*
- **Analyzing Process Data from Game/Scenario-Based Tasks: An Edit Distance Approach** by *Jiangang Hao, Zhan Shu and Alina von Davier*
- **Item Response Theory vs. Q-Matrices for Adaptive Testing** by *Jill-Jenn Vie, Fabrice Popineau, Jean-Bastien Grill, Eric Bruillard and Yolaine Bourda*
- **Evaluating Performance and Dropouts of Undergraduates using Educational Data Mining** by *Laci Mary Barbosa Manhaes, Sergio Manuel Serra Da Cruz and Geraldo Zimbrao*

- **Mining Large Scale Data from National Educational Achievement Tests** by *Reihaneh Rabbany, Osmar Zaiane and Samira Elatia*
- **Machine Learning with Crowdsourcing for Constructed Response Assessment: The Case of Free Speech Grading** by *Vinay Shashidhar, Nishant Pandey and Varun Aggarwal*
- **Towards Adaptive Education Assessments: Predicting Student Performance using Temporal Stability and Data Analytics in Learning Management Systems** by *Gautam S. Thakur, Mohammed Olama, Allen W. McNair, Sreenivas R. Sukumar and Scott Studham*

---

8. References

[1] Ambra Neri, Catia Cucchiarini, and Helmer Strik. Feedback in computer assisted pronunciation training: When technology meets pedagogy. *Proceedings of CALL professionals and the future of CALL research*, pages 179–188, 2002.

[2] Samuel Messick. Standards-based score interpretation: Establishing valid grounds for valid inferences. *ETS Research Report Series*, 1994(2):291–305, 1994.

[3] Amanda Pallais. Inefficient hiring in entry-level labor markets. Technical report, National Bureau of Economic Research, 2013.

[4] Dominic Barton Eric Labaye James Manyika Charles Roxburgh Susan Lund Siddarth Madhav Richard Dobbs, Anu Madgavkar. The world at work: Jobs, pay, and skills for 3.5 billion people, 2012.

[5] Stephen H Edwards. Using software testing to move students from trial-and-error to reflection-in-action. *ACM SIGCSE Bulletin*, 36(1):26–30, 2004.

[6] Sumit Gulwani, Ivan Radiček, and Florian Zuleger. Feedback generation for performance problems in introductory programming assignments. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*, pages 41–51. ACM, 2014.

[7] Kelly Rivers and Kenneth R Koedinger. Automatic generation of programming feedback: A data-driven approach. In *The First Workshop on AI-supported Education for Computer Science (AIEDCS 2013)*, page 50, 2013.

[8] Shashank Srikant and Varun Aggarwal. A system to grade computer programming skills using machine learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1887–1896. ACM, 2014.

[9] Vinay Shashidhar, Nishant Pandey, and Varun Aggarwal. Spoken english grading: Machine learning with crowd intelligence. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2089–2097. ACM, 2015.

[10] Aspiring Minds. National employability report: Engineering graduates, annual report 2011, 2011.

[11] Tiffany Barnes and John Stamper. Toward automatic hint generation for logic proof tutoring using historical student data. In *Intelligent Tutoring Systems*, pages 373–382. Springer, 2008.

[12] Lincoln C Wood, Hanna Teräs, and Torsten Reiners. The role of gamification and game-based learning in authentic assessment within virtual environments. 2013.

[13] Gráinne Conole and Bill Warburton. A review of computer-assisted assessment. *Research in learning technology*, 13(1), 2005.

[14] Pamela A Moss. Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62(3):229–258, 1992.

[15] Thomas K Landauer. Automatic essay assessment. *Assessment in education: Principles, policy & practice*, 10(3):295–308, 2003.

[16] S Barry Issenberg, William C McGaghie, Ian R Hart, Joan W Mayer, Joel M Felner, Emil R Petrusa, Robert A Waugh, Donald D Brown, Robert R Safford, Ira H Gessner, et al. Simulation technology for health care professional skills training and assessment. *Jama*, 282(9):861–866, 1999.

[17] Yossef S Ben-Porath and James N Butcher. Computers in personality assessment: A brief past, an ebullient present, and an expanding future. *Computers in Human Behavior*, 2(3):167–182, 1986.

[18] Udo Konradt, Guido Hertel, and Karin Joder. Web-based assessment of call center agents: development and validation of a computerized instrument. *International Journal of Selection and Assessment*, 11(2-3):184–193, 2003.

[19] John Hattie and Helen Timperley. The power of feedback. *Review of educational research*, 77(1):81–112, 2007.

[20] David A Waldman and Theresa Korbar. Student assessment center performance in the prediction of early career success. *Academy of Management Learning & Education*, 3(2):151–167, 2004.

[21] Ricardo Rosas, Miguel Nussbaum, Patricio Cumsille, Vladimir Marianov, Mónica Correa, Patricia Flores, Valeska Grau, Francisca Lagos, Ximena López, Verónica López, et al. Beyond nintendo: design and assessment of educational video games for first and second grade students. *Computers & Education*, 40(1):71–94, 2003.

[22] Francesco Bellotti, Bill Kapralos, Kiju Lee, Pablo Moreno-Ger, and Riccardo Berta. Assessment in and of serious games: an overview. *Advances in Human-Computer Interaction*, 2013:1, 2013.

[23] http://learningatscale.acm.org/las2016/.

[24] https://www.kaggle.com/c/the-allen-ai-science-challenge.