Incorporating Emotional Intelligence into Assessment Systems

Han-Hui Por and Aoife Cahill

Educational Testing Service, Princeton, NJ 08541, USA
{HPOR, ACAHILL}@ets.org

Abstract. This paper proposes developing emotionally intelligent assessments to increase score validity and reliability. We summarize research on three sources of data (process data, response data, and visual and sensory data) that identify students' needs during test taking and highlight the challenges in developing caring adaptive assessments. We conclude that because of its interdisciplinary nature, the development of caring assessments requires closer collaborations among researchers from diverse fields.

Keywords: Caring assessments, Adaptive testing, Emotions.

1 Introduction

Assessments can be stressful. Under-performance due to test anxiety can undermine the validity of a test score by failing to recognize the true performance of the student. Caring assessments [37] are systems developed to respond to students' needs by adjusting content sequencing, moderating the amount and type of feedback, adding visualization aids, etc. In addition to students' ability, caring assessment systems take into account additional information from both traditional and non-traditional sources (e.g., student emotions, prior knowledge and opportunities to learn). Caring assessments go beyond traditional assessments by providing the encouragement and resources that students might need.

One way to identify a student's learning needs is via their emotions during test-taking, such as the affective-sensitive version of AutoTutor [9], the uncertainty (i.e., confusion) adaptive version of ITSpoke (UNC-ITSpoke) [12], and emotion-sensitive versions of Cognitive Tutors and ASSISTments [1]. These systems have the "emotional intelligence" to recognize the emotions and needs of their learners and use the information to help the learners achieve their learning goals.

In this paper, we argue that a fair and valid caring assessment should take multiple interdisciplinary sources of information into account. We give an overview of current state-of-the-art capabilities -- both technological and psychometric -- that are relevant for designing and developing caring assessments. We outline some of the many challenges faced and suggest some areas for future work, particularly focusing on interdisciplinary collaborations. While we are particularly interested in the role of caring assessments in the context of summative assessments, parts of our discussion will also

refer to elements of caring assessments that will also be helpful for formative assessments.

2 Adaptive Testing

In caring assessments, students see different sets of questions and aids that are tailored to their ability and needs. In traditional computerized adaptive testing (CAT), pre-calibrated test items or testlets (sets of items, as in multi stage adaptive testing) are presented to students based on the quality of responses to previous questions. Given the response (correct/incorrect), an estimate of the student's ability is updated before further items are selected and presented. As a result of adaptive administration, different students experience different items. By successively fielding items selected to provide the maximum information about a student's ability, the maximum information about the student's ability is collected with each question, resulting in a shorter exam. Compared to a traditional static exam, a CAT exam is an assessment that tailors itself to each student's ability. Unfortunately, an algorithm that assigns items based on ability alone may not necessarily support a positive testing experience [27]. However, it may be possible to leverage psychometric approaches to incorporate additional information -- beyond ability -- into item selection in CAT.

2.1 Item Selection using Item Response Theory (IRT) and Bayesian Networks Models

IRT [19] is the current dominant statistical model used in CAT to assign students comparable, meaningful scores, even though students see sets of items tailored to their individual needs. Within the IRT framework, the estimated ability of a student is expected to be the same regardless of the items. The usefulness of IRT to select hints based on students' ability was demonstrated in FOSS (Full Option Science System), an ITS that was part of the NSF-funded Principled Assessment Designs for Inquiry (PADI) assessment [29].

To take into account both ability as well as the emotional needs of students, parameters of items in caring assessments can be estimated using models that integrate person- and variable-centered information such as the mixed-measurement item response theory (MM-IRT) [23, 24]. Such models focus on unobserved characteristics by identifying latent classes of individuals who respond to items in unexpected but distinct ways [13, 40]. Indeed, there is a recognition that observed scores may not correspond with unseen differences in how individuals respond to tests, such as differences in test strategies [23] or reactions to testing procedures [3].

Bayesian graph models from the Bayesian Network framework is another promising approach. These models allow the modeling of relevant conditional probabilities to update the likelihood of an event in the network. A major advantage is that the assessment of item mastery can be defined using multiple latent traits. In particular, it has been shown that the POKS (Partial Order Knowledge Structures) Bayesian modeling approach is computationally simpler and can outperform a 2-parameter IRT model in

some instances [7]. The Andes Tutor [5] and Hydrive [22] both incorporate Bayesian network models to select and score items.

3 Research on Emotions and Affective Computing

Caring assessments have to take into account students' emotional states (e.g., anxiety, frustration) during test-taking. Using self-reported measures, Lehman and Zapata-Rivera [18] identified the emotions that occurred when students completed conversation-based assessments (CBAs). They found that students experienced similar emotions across two studies and concluded that boredom, confusion, curiosity, delight, engagement/flow, frustration, happiness/enjoyment, hope, and pride are the prevalent emotions in CBAs. In an adaptive assessment, the use of self-reported measures disrupts test flow and other predictive indicators are necessary to identify emotions accurately. In addition to correct/ incorrect responses from the typical item types such as multiple choices items, we can enhance our prediction of students' emotions using three sources of data: process data, response data, and visual and sensory data.

Process Data such as response time, keystrokes, mouse clicks
Response Data such as responses to multiple choices or constructed response items, text or verbal feedback from students
Visual and Sensory Data such as eye tracking, heart rates, postures, facial expressions

The data and model complexity used in an assessment to predict emotions depends on the assessment objectives. Assessments used to identify areas of students' strengths, and weaknesses require detailed information about each student to provide very specific and individualized support. On the other hand, assessments conducted by teachers to inform teaching and learning direction require aggregated information of the group, and less precise information about individual students.

The amount of information also depends on the nature of the assistance the assessment aims to provide. A formative assessment is likely to provide more assistance in the form of additional visual aids, redirecting students' focus to the correct cues or paraphrasing questions when necessary. A summative assessment for determining proficiency levels can still benefit from collecting some amount of information to promptly identify students experiencing technical difficulties during the assessment. In the case of multi-year assessments, information can also be collected to aid in the development of exams in subsequent years. In the next section, we summarize research that has been done with each of the three information sources.

3.1 Process Data

Process data, such as typing speed, response time, keystrokes, mouse clicks and action sequences in problem solving tasks, trace students' progress through an assessment. Most process data can be collected in the background with minimal incremental costs

and is unobtrusive to students taking the exam. In general, they fall into three categories: what a student does, in what order, and how long it takes to do it. For instance, an analysis of the patterns and pauses in students' typing in a NAEP writing test showed that students who used the delete key more often, as a measure of their attempts to edit, had higher scores than students who did not delete as much [33]. The findings suggest that the latter group could benefit from encouragement to edit. Wise et al. [34] found that monitoring learners' response time and displaying warning messages when learners exhibited rapid guessing behavior improved scores and score validity, as indicated by the higher correlation between the test score and the learners' GPA and SAT scores. Other studies using timing data explored students' test taking behavior [15,16].

3.2 Response Data

Natural language processing (NLP) techniques are widely used to automatically measure the quality of constructed (free) responses in educational assessments. NLP is used in automated scoring engines to assess students' level of comprehension or writing proficiency and subsequently drive the feedback that students receive. Beigman-Klebanov et al. [2] show that by using NLP techniques it is possible to automatically predict a student's *utility value* -- a measure of how well the student can relate what they are writing about to themselves or other people -- from the student's writing. Flor et al. [10] show that it is possible to automatically categorize the dialogue acts (including expressing frustration) in a collaborative problem-solving framework using NLP techniques. NLP techniques are used on both written and spoken text. Spoken data can also provide a rich amount of information on both speaker emotions as well as their thought process (disfluencies, pause structure, etc.). Studies have also examined the use of low-level linguistic features to predict student emotions during human and computer tutoring sessions [17, 8]. Future research can focus on using complex linguistic analysis to learn more sophisticated relationships between the content of students' responses and their emotions in real time assessments.

3.3 Visual and Sensory Data

Visual and sensory data can also be captured to provide information on the students' progress or emotional state. The interest in sensory information stems from the findings that increased heart-rate and perspiration often precede our actual awareness of emotions, and studies have shown that heart rate and respiratory frequency can distinguish between neutral (relaxed), positive (joy) and negative (anger) emotions [31, 36]. However, while pulse rate monitors can be small, most devices would likely be obstructive when taking tests.

Although advances in facial recognition technology have vastly improved in recent years, identifying emotions accurately in real time is still a challenging task. Facial expressions are an integral part of emotions, but can also exist independently of emotions [25] and vice-versa. More recent developments suggest that new facial recognition algorithms have had some success with extracting features to classify students' emotions [35] in real time.

Eye tracking also allows us to pinpoint sections of the items that students are focused on. Studies on the usefulness of eye tracking data have provided preliminary evidence that they provide insights into how students respond to test items and solve problems [21, 28, 30]. A study that used eye tracking devices found that students with a history of performing poorly on reading tests did better when they had to write a summary of a reading passage before answering multiple-choice questions on the content [32]. The eye-tracking data showed that those students spent more time reading the initial text, and less time referencing the passage, suggesting that the students had built a mental memory model of the text. The advantage was stronger in students weaker in reading.

4 Challenges of Developing Caring Assessments

The development of caring assessments presents numerous challenges and research directions. Above all, the development of caring adaptive assessments requires closer interdisciplinary collaboration. Current research on the use of process data, NLP data, and visual/sensory data is largely focused on how these features correlate with either students' performance [21, 28, 30, 33, 34] or human raters in the field of automated scoring [11, 14, 20]. On the other hand, data from multiple sources would allow us to build accurate large-scale models of behavior from which we could then generalize students' behavior [4, 6, 38] and adapt to their needs. One area of future research is to focus on the predictive value of data from multiple sources in predicting students' emotions, and the impact of responding with aids on students' learning.

Other challenges include the costs and benefits of caring assessments over traditional ones. For caring assessments to be adopted as an industry standard, it will be necessary to demonstrate the effectiveness of the caring components both in terms of improving the student experience as well as contributing to overall test reliability and validity. As the approach to caring assessments is different in different educational context (e.g., summative vs formative), additional work is needed to define the elements that make up caring assessments so that the elements and combination of elements can be studied for their effectiveness.

In addition, the widespread adoption of caring assessments will be dependent on technology. Established assessments such as the GRE, TOEFL, LSAT depends on the capabilities of their testing centers. Therefore, if a caring assessment requires high-resolution cameras, all test centers would need to provide that hardware. In large-scale international assessments with test centers in all corners of the world, this is no small challenge.

4.1 Psychometrics Challenges in Caring Assessments

The psychometrics of caring assessments also presents some challenges. The current challenge is to adapt assessments based on students' ability and needs. While adaptive testing is not new, we need further research to establish if available models can accommodate multi modal, individual, and item level characteristics.

Scoring Complex Data Sequences

Another significant challenge for interactive assessments that respond to students' needs is that students can choose to take a large combination of actions. Should a student be rewarded with more points for taking fewer steps to get to the correct response? Recent research in psychometrics suggests that incorporating process data in assessments is tenable. A transition network using weighted directed networks can capture activity sequences, with nodes representing actions and directed links connecting two actions only if the first action is followed by the second action in the sequence [39]. As for scoring, Shu et al. [26] proposed a Markov-IRT model to characterize and capture the unique features of students' individual response process during a problem-solving activity in scenario-based tasks by laying out the model structure, its assumptions, the parameter estimation and parameter space. The Markov-IRT model allows test developers to determine the mapping of specific combinations to scoring rubrics.

Implications for Summative Assessments

Psychometric research can also contribute to scoring issues, particularly for high stakes summative assessments, where assigning valid and reliable scores that reflect students' skill mastery is a critical component. These assessments involve further issues such as score discrimination between students, in that students who score higher have better mastery than students with lower scores, and score comparability across cohorts of students who take different versions of the assessments.

Further, standard concerns in testing that are typically of lesser importance in learning assessments will surface. Issues such as fairness in testing, item overexposure, the establishing of cut scores, scaling and equating of scores, reporting and use of scores have been extensively studied and will also need to be adapted for a caring assessment.

5 Conclusion

We posit that caring assessments have a place in both formative and summative assessments. To get there, we will require that researchers from diverse backgrounds, such as computer science, engineering, natural language processing, learning, and psychometrics, work closely together to make sure that any new caring assessment is as valid and reliable as possible.

References

1. Baker, R.S., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V., Kusbit, G.W., Ocumpaugh, J., Rossi, L.: Towards sensor-free affect detection in cognitive tutor algebra. International Educational Data Mining Society (2012)
2. Beigman Klebanov, B., Burstein, J., Harackiewicz, J., Priniski, S., Mulholland, M.: Enhancing stem motivation through personal and communal values: Nlp for assessment of utility value in student writing. In: Proceedings of the 11th Workshop on Innovative Use of NLP

for Building Educational Applications. pp. 199-205. Association for Computational Linguistics, San Diego, CA (June 2016)
3. Bolt, D.M., Cohen, A.S., Wollack, J.A.: Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. Journal of Educational Measurement **39**(4), 331–348 (2002)
4. Calvo, R.A., D'Mello, S.: Affect detection: An interdisciplinary review of models, methods, and their applications. IEEE Transactions on affective computing **1**(1), 18-37 (2010)
5. Conati, C., Gertner, A., Vanlehn, K.: Using bayesian networks to manage uncertainty in student modeling. User modeling and user-adapted interaction 12(4), 371–417 (2002)
6. von Davier, A.A.: Computational psychometrics in support of collaborative educational assessments. Journal of Educational Measurement **54**(1), 3–11 (2017)
7. Desmarais, M.C., Pu, X.: A bayesian student model without hidden nodes and its comparison with item response theory. International Journal of Artificial Intelligence in Education **15**(4), 291–323 (2005)
8. D'Mello, S.K., Dowell, N., Graesser, A.C.: Cohesion relationships in tutorial dialogue as predictors of affective states. In: AIED. pp. 9–16 (2009)
9. DMello, S.K., Lehman, B., Graesser, A.: A motivationally supportive affect-sensitive autotutor. In: New perspectives on affect and learning technologies, pp. 113–126. Springer (2011)
10. Flor, M., Yoon, S.Y., Hao, J., Liu, L., von Davier, A.: Automated classification of collaborative problem solving interactions in simulated science tasks. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 31–41. Association for Computational Linguistics, San Diego, CA (June 2016)
11. Flor, M., Yoon, S.Y., Hao, J., Liu, L., von Davier, A.: Automated classification of collaborative problem solving interactions in simulated science tasks. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 31–41. Association for Computational Linguistics, San Diego, CA (June 2016)
12. Forbes-Riley, K., Litman, D.: Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Communication 53(9-10), 1115–1136 (2011)
13. Hernández, A., Drasgow, F., González-Romá, V.: Investigating the functioning of a middle category by means of a mixed-measurement model. Journal of Applied Psychology **89**(4), 687 (2004)
14. Jeon, J.H., Yoon, S.Y.: Acoustic feature-based non-scorable response detection for an automated speaking proficiency assessment. In: Thirteenth Annual Conference of the International Speech Communication Association (2012)
15. Lee, Y.H., Haberman, S.J.: Investigating test-taking behaviors using timing and process data. International Journal of Testing 16(3), 240–267 (2016)
16. Lee, Y.H., Jia, Y.: Using response time to investigate students' test-taking behaviors in a naep computer-based study. Large-scale Assessments in Education 2(1), 8 (2014)
17. Lehman, B., DMello, S.K.: Predicting student affect through textual features during expert tutoring sessions. Presented at the annual meeting of the Society for Text and Discourse (2010)
18. Lehman, B., Zapata-Riveria, D.: Student Emotions in Conversation-Based Assessments. IEEE Transactions on Learning Technologies (In Print)
19. Lord, F.: Application of item response theory to practical testing problems. Hillsdale, NJ, Lawrence Erlbaum Ass (1980)

20. Madnani, N., Cahill, A., Riordan, B.: Automatically scoring tests of proficiency in music instruction. In: Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications. pp. 217–222 (2016)
21. Mayer, R.E.: Unique contributions of eye-tracking research to the study of learning with graphics. Learning and instruction 20(2), 167–171 (2010)
22. Mislevy, R.J., Gitomer, D.H.: The role of probability-based inference in an intelligent tutoring system. ETS Research Report Series 1995(2) (1995)
23. Mislevy, R.J., Verhelst,N.: Modeling item responses when different subjects employ different solution strategies. ETS Research Report Series 1987(2) (1987)
24. Rost, J.: A logistic mixture distribution model for polychotomous item responses. British Journal of Mathematical and Statistical Psychology 44(1), 75–92 (1991)
25. Rozin, P., Cohen, A.B.: High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial ex- pressions of americans. Emotion 3(1), 68 (2003)
26. Shu, Z., Bergner, Y., Zhu, M., Hao, J., von Davier, A.A.: An Item Response Theory Analysis of Problem-Solving Processes in Scenario-Based Tasks. Psychological Test and Assessment Modeling 59(1), 109 (2017)
27. Shute, V.J., Hansen, E.G., Almond, R.G.: You can't fatten a hog by weighing it–or can you? evaluating an assessment for learning system called aced. International Journal of Artificial Intelligence in Education 18(4), 289–316 (2008)
28. Tai, R.H., Loehr, J.F., Brigham, F.J.: An exploration of the use of eye-gaze tracking to study problem-solving on standardized science assessments. International journal of research & method in education 29(2), 185–208 (2006)
29. Timms, M.J.: Using item response theory (IRT) to select hints in an ITS. Frontiers in Artificial Intelligence and Applications 158, 213 (2007)
30. Tsai, M.J., Hou, H.T., Lai, M.L., Liu, W.Y., Yang, F.Y.: Visual attention for solving multiple-choice science problem: An eye-tracking analysis. Computers & Education 58(1), 375–385 (2012)
31. Valderas, M.T., Bolea, J., Laguna, P., Vallverdú, M., Bailón, R.: Human emotion recognition using heart rate variability analysis with spectral bands based on respiration. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. pp. 6134–6137. IEEE (2015)
32. Wang, Z., Sabatini, J., OReilly, T., Feng, G.: How individual differences interact with task demands in text processing. Scientific Studies of Reading 21(2), 165–178 (2017)
33. White, S., Kim, Y.Y., Chen, J., Liu, F.: Performance of Fourth-Grade Students in the 2012 NAEP Computer-Based Writing Pilot Assessment: Scores, Text Length, and Use of Editing Tools. Working Paper Series. NCES 2015-119. National Center for Education Statistics (2015)
34. Wise, S.L., Bhola, D.S., Yang, S.T.: Taking the Time to Improve the Validity of Low-Stakes Tests: The Effort-Monitoring CBT. Educational Measurement: Issues and Practice 25(2), 21–30 (2006)
35. Yang, D., Alsadoon, A., Prasad, P., Singh, A., Elchouemi, A.: An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment. Procedia Computer Science 125, 2–10 (2018)
36. Yu, S.N., Chen, S.F.: Emotion state identification based on heart rate variability and genetic algorithm. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. pp. 538–541. IEEE (2015)
37. Zapata-Rivera, D.: Toward caring assessment systems. In: Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization. pp. 97–100. ACM (2017)

38. Zapata-Rivera, D., Liu, L., Chen, L., Hao, J., von Davier, A.A.: Assessing science inquiry skills in an immersive, conversation-based scenario. In: Big data and learning analytics in higher education, pp. 237–252. Springer (2017)
39. Zhu, M., Shu, Z., Davier, A.A.: Using Networks to Visualize and Analyze Process Data for Educational Assessment. Journal of Educational Measurement 53(2), 190–211 (2016)
40. Zickar, M.J., Gibby, R.E., Robie, C.: Uncovering faking samples in applicant, incumbent, and experimental data sets: An application of mixed-model item response theory. Organizational Research Methods 7(2), 168–190 (2004)