When Should an Adaptive Assessment Care?

Blair Lehman, Jesse R. Sparks, and Diego Zapata-Rivera

Educational Testing Service, Princeton NJ 08541, USA
[blehman, jsparks, dzapata]@ets.org

Abstract. Assessments can be a challenging experience for students. Students often have to consider more than just the knowledge being assessed, such as how to manage emotions that can impede performance (e.g., anxiety). But what if assessments cared about students and allowed them to just focus on the content of the assessment? In the present paper, we propose three time points at which assessments could care about students and discuss recent research that supports this model of assessments that care. The three time points include before, during, and after the assessment. Before students begin the assessment, the assessment format and design features can be adapted to the student; during the assessment adaptive support can be provided; and after the assessment students can be provided with personalized feedback. Adaptations would be made based on student characteristics (e.g., interest, self-efficacy) and behaviors during the assessment (e.g., emotions, response patterns). Ultimately, these adaptations at each time point would provide an individualized assessment experience for students, which could promote engagement and increase the quality of evidence collected about students' knowledge, skills, and abilities.

Keywords: Emotions, student characteristics, non-traditional assessments.

1 Introduction

Test taking has long been identified as an emotional experience for students (see Zeidner [1] for a review). Initially, research focused on test anxiety and its negative impact on performance in high-stakes assessments [1]. More recent research has also investigated the impact of students' low motivation or disengagement on performance in low-stakes assessments [2]. In both of these cases, the target emotion hinders students from performing to the best of their abilities on the assessment. Thus, students have an unpleasant experience and the assessment is not a valid measure of students' knowledge, skills, and abilities for those intending to use the scores.

Control-Value Theory [3], however, proposes a variety of positive (and negative) emotions that students are likely to experience during assessments and suggests that the positive emotions are beneficial for performance. Empirical research on overall student emotions for traditional assessments (e.g., multiple-choice items) has supported this proposed relationship [4]. Recent research on students' moment-to-moment emotions during a non-traditional assessment (e.g., conversation- and game-based assessments) has also shown that students experience both positive and

negative emotions and that engagement, specifically, is beneficial for performance [5].

The majority of research on emotions during assessments has focused on documenting the range of emotions that students experience. However, this information can also be leveraged to provide emotion-sensitive support to students. Emotion-sensitive support has been integrated into several intelligent tutoring systems (ITS; see [6] for a review). This type of support has been found to particularly benefit students that were struggling with the learning activity (e.g., [7], [8]). The result of integrating this type of support for students during assessments would be assessments that care [9]. These so-called "caring" assessments, which consider students' experience while completing the assessment, can benefit the student and improve assessment validity. Students can have a more positive experience while completing the assessment and the assessment can be used to gather more valid evidence of the students' knowledge, skills, and abilities because the student is more engaged with the task.

This type of on-demand emotion-sensitive support has only been explored in one computer-based assessment [10], although it has been more thoroughly investigated in the ASSISTments program [11] that blends tutoring and assessment (e.g., [12]) and in educational activities that aim to improve learning, as mentioned previously. The effort-monitoring computer-based assessment developed by Wise et al. [10] provided reminders to students when careless responding was detected and was successful at getting students to respond in a more effortful manner. It is also important to note that caring assessments should not be limited to responding only to student emotions; for example, research on ITSs has shown that behaviors such as gaming the system [13] and student characteristics such as domain-relevant interest and prior knowledge [14] impact students' experiences and learning outcomes.

In the present paper, we propose a model of caring assessments that includes three time points at which assessments can adapt: before, during, and after the assessment. The adaptive support provided by ITSs is usually limited to the time *during* the learning activity. In the context of assessment, we would like to propose expanding beyond the assessment activities themselves to include front-end selection (*before*) of both format and design features as well as end of assessment feedback (*after*). The adaptations would be based on student characteristics and behaviors observed during the assessment. This larger characterization of the assessment process is supported by recent research showing that students experience a variety of emotions during assessment preparation (i.e., studying), assessment completion, and review of assessment performance feedback [15]. Next, we will discuss our proposed model for assessments that care.

2 Assessments that Care

We propose that caring can be integrated into assessments at three time points (before, during, and after the assessment) through various types of adaptations based on student characteristics and behaviors. Student characteristics can include more general

personality traits as well as beliefs and perceptions about a specific domain. Student behaviors are dependent upon the attributes and content of the assessment and include the actions students take within the environment. These actions can be used to infer cognitive, emotional, and motivational states. The adaptations that can be made based on these inferences are often dependent on decisions made at previous time points. For example, on-demand adaptations to respond to student disengagement (e.g., providing motivational statements) will be constrained based on the previously selected assessment format (e.g., conversation-based assessment vs. traditional assessment). Thus, it is important to consider how adaptations build upon each other to create a more engaging assessment experience. Next, we discuss each time point and research that suggests that these adaptations are advantageous for students.

2.1 Time 1: Before the Assessment

The first time point is before the student begins the assessment. At Time 1 there are two types of adaptations that can occur: adaptations to the assessment format or to the design features of the assessment. Both types of adaptation would require information about student characteristics prior to administration of the assessment to create a student profile that would be used for adaptation decisions. Thus, before the assessment can be administered, information would need to be collected from students. This could potentially be problematic as the collection of additional information could either increase the total time for a test administration or require a separate administration session. Next, we will discuss each adaptation separately.

The main decision for assessment format is whether to have students complete a traditional assessment, a non-traditional assessment, or an assessment that has both types of items. Non-traditional assessments have been developed, in part, because they are hypothesized to provide a more engaging experience for students. This more engaging experience is proposed to then result in students performing to the best of their abilities. Recent research on game-based assessments (GBA) has shown that student performance is typically positively correlated with a more positive experience (e.g., [16]). However, there have been very few efforts directly comparing performance and experience between different assessment formats that assess the same knowledge and skills.

One exception comes from research on GBAs that assess argumentation skills. Lehman, Jackson, and Forsyth [17] compared student performance and experience on a traditional assessment and a GBA. The findings revealed that students who performed better on one assessment format than the other reported different emotional experiences. Specifically, students that performed better on the GBA compared to the traditional assessment reported more positive experiences during the GBA than those who performed worse on the GBA. However, this work did not explore the student characteristics that could be predictive of which assessment format afforded students the opportunity to perform to the best of their ability and have a positive experience. Knowledge of the relevant student characteristics would be critical to enable effective *a priori* assignment of students to a particular assessment format.

After the assessment format has been selected, the next opportunity for adaptation is what version of the assessment to administer to the student. By version, we mean that there is more than one option for the design of the tasks within the same assessment format assessing the same knowledge and skills. These different versions may involve varying more superficial aspects of the environment (e.g., surface features or presentation mode) to accommodations for students with disabilities. It is likely that non-traditional assessments will afford more opportunities for a variety of versions as they often include more elements to the environment such as agents who can have different characteristics or assume different roles. These design features can then be adapted to meet the students' needs.

Sparks, Zapata-Rivera, Lehman, James, and Steinberg [18] have begun investigating the use of different assessment versions in the context of a conversation-based assessment (CBA) that assesses science inquiry skills. Four versions of the CBA were developed that varied the knowledge level of a virtual peer agent (high, low) and how questions were framed (comparison, agreement). The findings revealed that overall the type of assessment evidence that could be collected varied for each version of the CBA and that the CBA version interacted with student characteristics (urban vs. rural school, prior knowledge). These findings suggest that some students could benefit more from different combinations of assessment design features rather than presenting all students with the same version of the assessment. It is also important to note that both types of adaptations before the assessment will require careful evaluation of the validity and equating of different assessment formats and versions to ensure comparability of scores across assessments (discussed further below).

2.2 Time 2: During the Assessment

The second time point at which adaptations can be employed to care about students is during the assessment. This type of adaptation is similar to the type of support that students receive from ITSs designed for learning. Specifically, there would be two layers of adaptation that encompass the inner and outer loop that dynamically select reactions to students' immediate actions (e.g., type of feedback) (inner loop) and adaptively select the next task for students to complete (outer loop) [19]. These adaptations can also include supports that address students' cognitive, emotional, and motivational states. Regardless of the type of support, these are all deployed based on an underlying student model that tracks students' knowledge and other states (e.g., gaming the system [13]) based on their behaviors in the environment.

Although a student model that includes information about students' cognitive, emotional, and motivational states has been incorporated into ASSISTments (e.g., [11], [12]), pure assessments (i.e., where learning is not an explicit goal) have generally utilized a less well-developed student model. Typically, computer adaptive assessments only include cognitive states (i.e., response quality as an indicator of knowledge level). One exception comes from the previously mentioned Wise et al. [10] study in which adaptive motivational support was successfully provided when student effort was monitored through response times. Recent research on student

emotions during CBAs has revealed instances in which emotion-sensitive support could be beneficial for students [5]. For example, high intensity frustration was found to be persistent, grow in frequency over time, and be negatively related to performance. This finding suggests that a more complex student model that includes cognitive, emotional, and motivational states could benefit students during assessments.

2.3 Time 3: After the Assessment

The third time point at which assessments can care is when students receive feedback about the quality of their performance on the assessment. We have included performance feedback as part of the assessment process because its perceived utility is important for assessment validity [20]. Specifically, if students receive feedback that is difficult to understand, vague about how to make improvements, or demotivating, then the assessment is not effective as a tool for improving students' knowledge as students will be less likely to engage in productive learning behaviors after the assessment. It is important to note that feedback could potentially occur during the assessment as well. Given that feedback during the assessment is not always appropriate or desirable, we have chosen to only focus on feedback provided after the assessment. However, Time 2 could be expanded to incorporate the use of feedback, particularly in the case of formative assessments where such feedback may be more appropriate.

Score reports are often used to provide information about performance after an assessment, and the majority of score reporting research has focused on how to clearly display information such as measurement error [21]. However, some researchers have proposed that score reports should differ by audience (e.g., students vs. teachers) [20] and should be increasingly interactive [22]. We propose to go a step further when taking the audience into consideration. Specifically, we would like score reports to be individually tailored to each student. The individualized score reports would utilize the student model (student characteristics and behaviors) from the assessment to provide contextualized information about the quality of performance and practical next steps to improve performance [23]. Importantly, this report would need to be presented in a way that is meaningful to students and motivates them to engage in the strategies to improve future performance. We view the presentation of this tailored report as particularly important because if students do not view the report as useful, or are unwilling to adapt their future behaviors based on the report, the accuracy of the score report itself becomes less important.

3 Conclusion

We have proposed three time points at which adaption could be incorporated into assessment development to create "caring assessments." The three time points we proposed include *before* students begin the assessment (assessment format and design features), *during* the assessment (on-demand support), and *after* the assessment (personalized feedback). We have expanded the opportunities for caring beyond the

assessment itself to encompass adaptations based on student characteristics outside of the assessment and the presentation of feedback after an assessment has been completed. However, it may be necessary to also include support for assessment preparation (i.e., studying) to create a complete caring assessment package [15].

Systems that provide adaptive support based on students' behaviors and even students' emotions during an educational activity are nothing new. There have been a variety of ITSs that detect, track, and respond to student emotions (see [6] for a review). However, this type of adaptivity has rarely been employed in educational activities that have assessment as the primary or only goal. There are two potential reasons for not including this type of adaptation in assessments. First, any type of adaptation will create a different assessment experience for students, which can make it more difficult for students' performance on the assessment to be equated. As mentioned previously, asking students to complete different assessments (formats and/or design features) requires that all of the assessments be equated to ensure that performance outcomes are comparable across the assessments. Equating is already part of assessment development when different forms of the same assessment are created [24], but the type of dynamic support that would be provided in an assessment that cares would likely further complicate the equating process.

The second reason that adaptive support has been employed more frequently in learning than in assessment activities has to do with the type of support that can be provided. An adaptive system that has the goal of facilitating student learning can provide a variety of support that gradually leads students towards the correct answer, or even provides the correct answer when students are struggling. This type of support is not likely to be useful when the goal of the system is to accurately assess students' current level of understanding. However, this does not mean that other types of adaptive support could not be utilized. For example, *Affect-Sensitive AutoTutor* [7] employs an intervention that targets both students' attributions and motivation. When students are found to be bored, confused, or frustrated, the tutor agent states that the students' current negative emotion was due to either the nature of the material (e.g., "This material is really challenging") or to the tutor (e.g., "I probably didn't explain the information very well") (attribution), followed by a statement encouraging the student to persist with the learning session (motivation). A similar approach could be adopted in assessments when students become disengaged; however, research is needed to determine the most effective approaches based on the student and the context.

We have presented some initial evidence that supports our proposed model of caring assessments. However, the evidence that we have presented is limited, in many cases to one study or context, and is only correlational. Thus, there are two critical next steps for future research in the development of caring assessments. First, the student characteristics that are most relevant to each time point of adaptation need to be identified. Second, the model needs to be tested for effectiveness of adaptations at each of the individual time points and for the overall model. It is important that we understand not only how adaptations at each time point impact students' performance and experience, but also how the adaptations interact across time points to impact the assessment. These caring assessments are hypothesized to provide three advantages:

(1) students will be more engaged and more likely to perform to the best of their ability, which in turn (2) will allow the assessment to collect more valid evidence of students' knowledge, skills, and abilities, and (3) students' more positive assessment experience may lead to more positive feelings in general about the domain and help to build students' self-efficacy. In other words, caring assessments will benefit a wide range of stakeholders who are involved in the assessment process.

References

1. Zeidner, M.: Test anxiety: The state of the art. Plenum Press, New York (1998).
2. Wise, S., Smith, L.: The validity of assessment when students' don't give good effort. In: Brown, G., Harris, L. (eds.) Handbook of Human and Social Conditions in Assessment, pp. 204-220. Routledge, New York (2016).
3. Pekrun, R.: The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. Educational Psychology Review, 18, 315-341 (2006).
4. Pekrun, R., Goetz, T., Frenzel, A., Barchfield, P., Perry, R.: Measuring emotions in students' learning and performance: The Achievement Emotions Questionnaire (AEQ). Contemporary Educational Psychology, 36, 36-48 (2011).
5. Lehman, B., Zapata-Rivera, D.: Intensity is as important as frequency for emotions during test taking. Contemporary Educational Psychology (in preparation).
6. Sottilaire, R., Graesser, A., Hu, X., Goldberg, B. (eds.): Design recommendations for intelligent tutoring systems, Vol. 2, Instructional Management. U.S. Army Research Laboratory, Orlando (2014).
7. D'Mello, S., Lehman, B., Graesser A.: A motivationally supportive affect-sensitive AutoTutor. In: Calvo, R., D'Mello, S. (eds.) New perspectives on affect and learning technologies, pp. 113-126. Springer, New York (2011).
8. Forbes-Riley, K., Litman, D.: Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. Speech Communication, 53, 1115-1136 (2011).
9. Zapata-Rivera, D.: Toward caring assessment systems. In: Tkalcic, M., Thakker, D., Germanakos, P., Yacef, K., Paris, C., Santos, O. (eds.) Adjunct Proceedings of User Modeling, Adaptation and Personalization Conference, pp. 97-100. ACM, New York (2017).
10. Wise, S., Bhola, D., Yang, S.-T.: Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. Educational Measurement: Issues and Practice, 25, 21-30 (2006).
11. Razzaq, L., Feng, M., Nuzzo-Jones, G., Heffernan, N., Koedinger, K., Junker, B., et al.: The Assistment Project: Blending assessment and assisting. In: Looi, C., McCalla, G., Bredeweg, B., Breuker, J. (eds.) Proceedings of the Artificial Intelligence in Education Conference, pp. 555-562. ISO Press, Amsterdam (2005).
12. Pardos, Z., Baker, R., San Pedro, M., Gowda, S., Gowda, S.: Affective states and state tests: Investigating how affect and engagement during the school year predict end of year learning outcomes. Journal of Learning Analytics, 1, 107-128 (2014).
13. Baker, R., Corbett, A., Roll, I., Koedinger, K.: Developing a generalizable detector of when students game the system. User Modeling & User Adapted Interaction, 18, 287-314 (2008).

14. Lehman, B., D'Mello, S., Graesser, A.: Who benefits from confusion during learning? An individual differences cluster analysis. In: Yacef, K., Lane, C., Mostow, J., Pavlik, P., (eds.) Proceedings of the Artificial Intelligence in Education Conference, pp. 51-60. Springer-Verlag, Berlin/Heidelberg (2013).
15. Peterson, E., Brown, G., Jun, M.: Achievement emotions in higher education: A diary study exploring emotions across an assessment event. Contemporary Educational Psychology, 42, 82-96 (2015).
16. Jackson, G., Lehman, B., Forsyth, C., Grace, L.: Game-based assessments: Investigating relations between skill assessment, game performance, and user experience. Computers in Human Behavior (in review).
17. Lehman, B., Jackson, G., Forsyth, C.: A (mis)match analysis: Examining the alignment between test-taker performance in conventional and game-based assessments. Journal of Applied Testing Technology (in preparation).
18. Sparks, J.R., Zapata-Rivera, D., Lehman, B., James, K., Steinberg, J.: Simulated dialogues with virtual agents: Effects of agent features in conversation-based assessments. In: Proceedings of Artificial Intelligence in Education Conference (2018).
19. VanLehn, K.: The behavior of tutoring systems. International Journal of Artificial Intelligence in Education, 16, 227-265 (2006).
20. Zapata-Rivera, D., Katz, I.: Keeping your audience in mind: Applying audience analysis to the design of interactive score reports. Assessment in Education: Principles, Policy, and Practice, 21, 442-463 (2014).
21. Zwick, R., Zapata-Rivera, D., Hegarty, M.: Comparing graphical and verbal representations of measurement error in test score reports. Educational Assessment, 19, 116-138 (2014).
22. Zapata-Rivera, D., Underwood, J., Bauer, M.: Advanced reporting systems in assessment environments. In: Kay, J., Lum, A., Zapata-Rivera, D. (eds.) Proceedings of Learner Modeling for Reflection Workshop at the Artificial Intelligence in Education Conference, pp. 23-31 (2005).
23. Zapata-Rivera, D.: Adaptive score reports. In: Masthoff, J., Mobasher, B., Desmarais, M., Nkambou, R. (eds.) Proceedings of the User Modeling, Adaptation, and Personalization Conference, pp. 340-345. Springer, Berlin/Heidelberg (2012).
24. Kolen, M., Brennan, R.: Test equating, scaling, and linking: Methods and practices (2nd Ed.). Springer, New York (2004).