Assessments That Care About Student Learning

Stephen E. Fancsali & Steven Ritter

Carnegie Learning, Inc., Pittsburgh PA 15219, USA
{sfancsali, sritter}@carnegielearning.com

Abstract. We argue that an important requirement of assessments that care is that they focus on student learning. Intelligent tutoring systems (ITSs) are a basis for such assessments; they provide a means by which to continually assess what students know *as they learn*. Given widespread dissatisfaction with high-stakes assessments, we present a review of recent work targeted at replacing high-stakes exams with regular use of an ITS. We conclude by discussing some areas for future research and development.

Keywords: Intelligent Tutoring Systems, Mathematics Education, High-Stakes Testing, Formative Assessment, Summative Assessment, Instruction-Embedded Assessment.

1 Introduction

1.1 Characteristics of Assessments that "Care"

John Self's [1] description of ITSs as systems that "care" about students focused on the way that the personalization in such systems allows them to care about students in a way that other systems cannot. With respect to caring assessments, we agree with Zapata-Rivera [2] that personalization can enable assessments to address students at their individual level of understanding. Personalization in caring assessments might also enable students to demonstrate their knowledge in different ways and, perhaps, at different times. However, the most important characteristic of a caring assessment is not a result of personalization but of the goal of the assessment. For an assessment to be "caring," the experience must be beneficial to the student. Summative assessments are typically, though not always, designed to benefit institutions by providing them with information about the effectiveness of some aspect of instruction (e.g., the teacher, institution, or materials). Students are merely measurement instruments in this process. In contrast, caring assessments are fundamentally formative and directly assist the students in learning.

We posit that an exciting opportunity exists wherein ITSs, augmented by several tools and affordances that still need to be developed, are used *as caring assessments*. Such assessments are fundamentally formative, focused on student learning, and adaptive to student differences, but they also can serve a summative purpose to the institution.

In what follows, we argue that the time has come, both technologically and politically, to push forward with innovative approaches to assessment that use technologies like ITSs, embedded within the learning process, to provide continual, on-going, formative assessment *while students learn* to replace high-stakes, end-of-year summative assessment approaches. Accomplishing this goal, relying on systems like ITSs that attend to Self's notion of "caring" about students (e.g., by having a student model of what learners know and do not know *during the learning process*), will better allow a broad swath of educators, courseware and ITS developers, and others to (eventually) bask "in the positive glow associated with the term" caring [1]. More importantly, innovative approaches will increase instructional time, provide better measures of what students actually know, and improve learning outcomes. We detail recent work in developing statistical models that predict students' end-of-year test scores in mathematics using data from an "ITS that cares," namely Carnegie Learning's MATHia ITS, based on its Cognitive Tutor technology [3].

1.2 The Problem(s) with High-Stakes, Summative Assessments

High-stakes summative assessments, by design and implementation, often contradict what we know to be beneficial to instruction [4]. The fact that only the student's knowledge on the particular day of the test is important leads to cramming, which optimizes short-term performance, at the expense of long-term memory [5,6]. Item Response Theory (IRT) assumes that student knowledge is fixed for the period of the exam, and so the examination environment is set up to minimize student learning (even though we do know that prompted memory retrieval, as practiced in tests, does improve learning [7]).

Most high-stakes assessments only provide coarse measures of learning like multiple-choice items, which, even when well designed (e.g., with demonstrated validity and reliability), provide minimal opportunities to illuminate student misconceptions or the extent to which learners have mastered particular micro-competencies, skills, or knowledge components (KCs [8]).

In addition to the aforementioned shortcomings, standardized, high-stakes, summative assessments crowd out instructional time. Not only does taking the tests take time, but also teachers often spend several instructional periods (and in many cases weeks' worth of instructional time) preparing for such high-stakes assessments. Further, there are often numerous tests given. The Council of Great City Schools reports that, among large school districts recently surveyed in the U.S., the typical eighth grader, in a typical academic year, spends 25.3 hours taking 10.3 *district-administered* tests, which alone would consume 2% of instructional time in a 180 instructional-day academic year, without accounting for preparation time and other summative assessments [9].

1.3 Responses to the Problem(s)

Public backlash to perceived and actual shortcomings of high-stakes, standardized testing reflects perceptions that testing takes up too much instructional time while not being well-aligned to such instruction [10]. On a national level in the U.S., the Every

Student Succeeds Act (ESSA) encourages innovative assessment approaches, demonstrating recognition that the existing framework is less than satisfactory. At a state and local level, so-called "opt-out" movements [11] have led to parents and students exercising their rights to not be required to take certain high-stakes, standardized assessments. As we noted in [12], in 2017, 27% of students in the U.S. state of New York opted out of high-stakes math testing [13], and so many students in Minneapolis recently opted-out of state exams for 10th and 11th grade math that the state does not believe that the exam results can be judged to be reliable [14]. Officials and legislators in Georgia (and elsewhere) are presently working to pursue possible alternatives to high-stakes, end-of-year assessments via possibilities like more frequent, formative assessments via short quizzes and other possible alternatives [15]. What these response tend to have in common is a recognition that accountability and assessment of learning and knowledge are important but that the methods presently employed to assess such learning and knowledge are inadequate.

1.4 MATHia & Cognitive Tutor

MATHia is an ITS for middle school and high school mathematics, based on Carnegie Learning's Cognitive Tutor technology, that typically is a part of a blended mathematics curriculum. Carnegie Learning generally recommends that the instructional mix of this blended curriculum be a 60%-40% split between instructor-facilitated, student-centered classroom activities that facilitate collaborative learning and deep conceptual understanding (60% of the time) and individual student work in a computer lab or classroom with the MATHia ITS (40% of the time).

MATHia is based on an adaptive, mastery learning [16] approach and relies on a fine-grained model of KCs (e.g., Grade 6 mathematics comprised of approximately 700 KCs) that students must master to make progress through content. Content is presented to students in topical "workspaces," each of which focuses on a set of KCs that must be mastered to move on to the next workspace. Within each workspace, students work on multi-step, complex, real-world problems (see Fig. 1), and student responses at each step provide rich data about student problem-solving strategies and a fine-grained understanding of what students know and do not know.

2 Using MATHia Data to Predict Standardized Test Scores

Recent efforts [12, 17] have focused on using student MATHia performance data to predict standardized test scores in large school districts in the U.S. states of Virginia (VA) and Florida (FL). This work follows in the tradition of work using data from the ASSISTments system [18] and considers the relative contributions of various measures of MATHia performance (and transformations thereof) (e.g., workspaces mastered per hour, hints requested, errors made), prior year test performance or a pre-test score (i.e., prior knowledge), and socio-demographic data (e.g., socio-economic status via free/reduced-price lunch status, English language learner status, etc.).

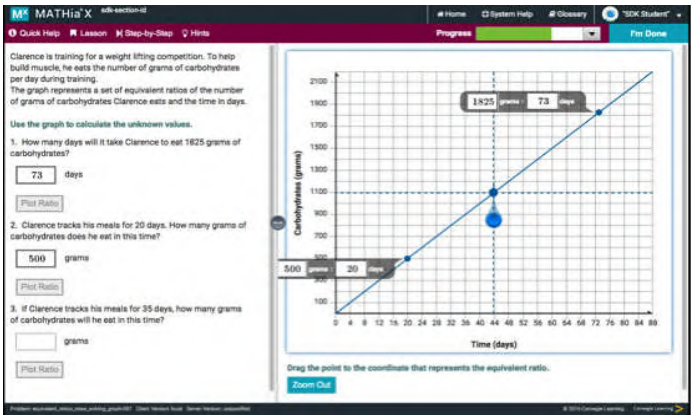

Fig. 1. A screenshot of problem-solving in the MATHia platform.

Specifics of model construction, specification, and selection are beyond the scope of the present discussion (see [12, 17]), but Table 1 provides a brief summary of results to demonstrate our success so far. While various model goodness-of-fit metrics are considered in detail in the original work reporting these results, we rely on the relatively simple to interpret adjusted R^2 values of the best models for particular academic years in Table 1.

In FL, the Florida Comprehensive Assessment Test (FCAT) was used in 2013-14, and the Florida Standards Assessment (FSA) was used in 2014-15 and 2015-16. Results for FL are reported are for the best model learned on data from another academic year's data, so in each case, results reported are for the situation in which an academic year's data served as a held-out test set for the statistical model learned [12]. In VA, models were learned to predict scores on the Standards of Learning (SOL) exam for mathematics [17], but data were only available for a single academic year. R^2 values reflect the proportion of variance in SOL exam scores explained by a model learned on data for 7th graders. Cross-validation results indicated that these values do not seem to reflect substantial over-fitting.

Table 1 shows that we can account for up to 73% of the variation in FSA scores, and we see the relative contribution of different categories of variables, starting with a model including pre-test scores (M1) and progressively increasing the complexity of models through M5. Importantly, we see that there are relatively small differences between M5 and M6 (which does not include demographics), so demographic variables do not provide for substantial predictive power. Ideally, we would be able to rely on process variables (i.e., MATHia performance) only, and especially for predicting FCAT/FSA, we explain over 50% of the variation in these scores with process/performance data alone.

Table 1. Adjusted R^2 values for best linear regression models reported in [12, 17]. Variable categories are pre-test performance (*pre-test*), MATHia process data (*process*), and demographic data (*demog*). An M6 model was not considered by [17].

Model	Variables	VA SOL	FL FCAT/FSA		
		2011-12	***2013-14***	***2014-15***	***2015-16***
		n=940	n=7,491	n=7,368	n=8,065
M1	*pre-test*	.5	.6001	.6035	.6528
M2	*process*	.43	.5271	.5393	.593
M3	*process* + *demog*	.45	.5443	.5656	.6185
M4	*pre-test* + *demog*	.51	.6059	.629	.6684
M5	*pre-test* + *demog* + *process*	.57	.6642	.689	.7349
M6	*pre-test* + *process*		.6707	.6326	.7258

3 Future R&D

Being able to predict standardized test scores with reasonable success using performance data from systems like MATHia is insufficient for such systems to replace such tests. Further, systems like MATHia are designed to be used and generally, though not exclusively, are used as a part of a blended curriculum. To transition to using such systems in an assessment role, we see several important areas of R&D to pursue both for Carnegie Learning and the broader community of ITS and assessment researchers working on developing caring assessments. In addition to improving models like those for which we have here briefly reported results, we need to identify minimally sufficient sets of content that contribute to successful predictive models. This will help to identify subsets of content that should be used as a part of assessments in ITSs like MATHia. Content management, assessment design, and editing tools will be required to allow for state-by-state and possibly local customization. Security tools will be required to insure that students do their own work. More work needs to be done to establish the validity and reliability of this approach to assessment, likely by continuing to build bridges between traditional IRT approaches and the knowledge tracing approaches of systems like MATHia.

References

1. Self, J.A.: The distinctive characteristics of intelligent tutoring systems research: ITSs care, precisely. International Journal of Artificial Intelligence in Education 10, 350–364 (1999).
2. Zapata-Rivera, D.: Toward caring assessment systems. In: Tkalcic, M, Thakker, D., Germanakos, P., Yacef, K., Paris, C., Santos, O. (eds.) Adjunct Publication of the 25th Conf. on User Modeling, Adaptation and Personalization, UMAP '17, pp. 97–100. ACM, New York (2017).
3. Ritter, S., Anderson, J.R., Koedinger, K.R., Corbett, A.T.: Cognitive Tutor: applied research in mathematics education. Psychonomic Bulletin & Review 14, 249–255 (2007).

4. Snow, R.E., Lohman, D.F.: Implications of cognitive psychology for educational measurement. In: Linn, R.L. (ed.) Educational Measurement, 3rd ed., pp. 263–331. American Council on Education/Macmillan, New York (1989).
5. Bloom, K.C., Shuell, T.J.: Effects of massed and distributed practice on the learning and retention of second-language vocabulary. Journal of Educational Research 74(4), 245–248 (1981).
6. Rea, C.P., Modigliani, V.: The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. Human Learning: Journal of Practical Research & Applications 4(1), 11–18 (1985).
7. Roediger, H.L., Karpicke, J.D.: The power of testing memory: Basic research and implications for educational practice. Perspectives on Psychological Science 1, 181–210 (2006).
8. Koedinger, K.R., Corbett, A.T., Perfetti, C.: The Knowledge-Learning-Instruction (KLI) framework: Bridging the science-practice chasm to enhance robust student learning. Cognitive Science 36(5), 757–798 (2012).
9. Hart, R., Casserly, M., Uzzell, R., Palacios, M., Corcoran, A., Spurgeon, A.: Student testing in America's great city schools: An inventory and preliminary analysis. Council of Great City Schools, Washington, DC (2015).
10. PDK/Gallup.: 47th annual PDK/Gallup poll of the public's attitudes toward the public schools: Testing doesn't measure up for Americans. Phi Delta Kappan 97(1), (2015).
11. Bennett, R.E.: Opt out: An examination of issues. ETS Research Report No. RR-16-13 (ETS Research Report Series). Educational Testing Service, Princeton, NJ (2016) doi:10.1002/ets2.12101
12. Fancsali, S.E., Zheng, G., Tan, Y., Ritter, S., Berman, S.R., Galyardt, A.: Using embedded formative assessment to predict state summative test scores. In: Proceedings of the 8th International Conf. on Learning Analytics and Knowledge, pp. 161–170. ACM, New York (2018).
13. Moses, S.: State testing starts today; opt out CNY leader says changes are 'smoke and mirrors.' Syracuse.com (28 March 2017). http://www.syracuse.com/schools/index.ssf/2017/03/opt-out_movement_ny_teacher_union_supports_parents_right_to_refuse_state_tests.html, last accessed 2018/03/29.
14. State of Minnesota, Office of the Legislative Auditor: Standardized student testing: 2017 evaluation report. State of Minnesota, Office of the Legislative Auditor, St. Paul, MN (2017).
15. Tagami, T.: Smaller tests could replace state's big Milestones exams. The Atlanta Journal-Constitution (02 February 2018). https://www.myajc.com/news/local-education/smaller-tests-could-replace-state-big-milestones-exams/xbdXop4VvI2Tmf6EFl7fVN/, last accessed 2018/03/29.
16. Bloom, B.S.: Learning for mastery. Evaluation Comment 1(2), (1968).
17. Ritter, S., Joshi, A., Fancsali, S.E., Nixon, T.: Predicting Standardized Test Scores from Cognitive Tutor Interactions. In: Proceedings of the Sixth International Conference on Educational Data Mining, pp. 169–176. (2013).
18. Junker, B.W.: Using on-line tutoring records to predict end-of-year exam scores: experience with the ASSISTments project and MCAS 8th grade mathematics. In: Lissitz, R.W. (ed.) Assessing and modeling cognitive development in school: intellectual growth and standard settings. JAM, Maple Grove, MN (2006).