Assessing Writing 21 (2014) 104–111

Contents lists available at ScienceDirect

Assessing Writing

Forum

When "the state of the art" is counting words

Les Perelman*

Massachusetts Institute of Technology, United States

ARTICLE INFO

Article history:
Received 7 May 2014
Accepted 19 May 2014
Available online 12 June 2014

Keywords:
Automated essay scoring
Common Core standard
Essay length
High-stakes assessment
Race-to-the-top
Human raters

ABSTRACT

The recent article in this journal "State-of-the-art automated essay scoring: Competition results and future directions from a United States demonstration" by Shermis ends with the claims: "Automated essay scoring appears to have developed to the point where it can consistently replicate the resolved scores of human raters in high-stakes assessment. While the average performance of vendors does not always match the performance of human raters, the results of the top two to three vendors was consistently good and occasionally exceeded human rating performance." These claims are not supported by the data in the study, while the study's raw data provide clear and irrefutable evidence that Automated Essay Scoring engines grossly and consistently over-privilege essay length in computing student writing scores. The state-of-the-art referred to in the title of the article is, largely, simply counting words.

Much of the enthusiasm for using automated essay scoring is motivated by the increased number of writing assessments informed by the Common Core standards and mandated by the U. S. Department of Education's Race-to-the-Top initiative. The stakes for getting these assessments right are very high for students, teachers, schools, school districts, and states. States are compelled by the No Child Left Behind law to use standardized test scores in teacher evaluations for tenure, pay, and promotion, as evidenced by the severe economic sanctions the Federal government has recently placed on State of Washington (Higgins, 2014). Consequently, it is inevitable that assessment will, to a large extent, define instruction. The two major Race-to-the-Top Consortia, Partnership for Assessment of Readiness for College and Careers (PARCC) and SMARTER BALANCED Assessment Consortium, are under intense

* Address: Room 14E-403, Massachusetts Institute of Technology, Cambridge, MA 02139. Tel.: +017818623833.
E-mail address: perelman@mit.edu

http://dx.doi.org/10.1016/j.asw.2014.05.001
1075-2935/

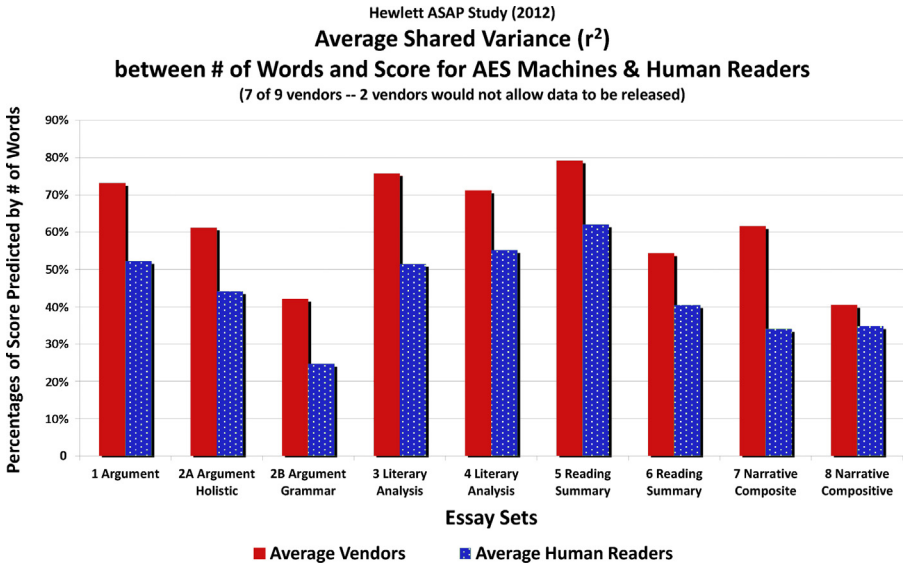

Fig. 1. Average shared variance between # of words and scores for human readers and AES machines.
Source: Calculations derived from data obtained at Automated Student Assessment Prize (2013).

pressure to cut costs. Indeed, ten of the original twenty-six PARCC states have withdrawn from the consortium largely because of cost, leaving only sixteen states and the District of Columbia (Ujifusa, 2014). At the same time, Automated Essay Scoring (AES) presents a huge economic advantage to testing companies by potentially reducing the marginal cost of scoring essays to close to zero.

It is no wonder, then, that there were large incentives to conduct the ASAP competition and to believe Professor Shermis' assertion in his article in this journal that "Automated essay scoring appears to have developed to the point where it can consistently replicate the resolved scores of human raters in high-stakes assessment" (Shermis, 2014, p. 75). Unfortunately, the data provided in that article and in the link to the raw data provided do not substantiate this claim.

The following analysis derives from three sources: the summary data presented in that article and two earlier versions of it (Shermis & Hamner, 2012, 2013), the training data downloaded from the Kaggle competition site (Kaggle, 2012), and the incomplete set of raw data from the ASAP site http://www.scoreright.org/asap.aspx?content=Request_ASAP_Phase_One_Data.

Of the nine named vendors in the study, two refused permission to have their data released. Moreover, although all participating vendors were identified in Shermis, 2014, the released raw data was anonymous, with vendors being identified only as Vendor1, Vendor2, etc. Furthermore, one of the conditions of the Terms-of-Service in downloading the data, was to refrain from any attempt to identify the participating vendors. The figure and two tables in this study are derived from my analyses of these raw data.

The principal value of Professor Shermis' study, although probably unintentional, is that the raw data of the study provide clear and irrefutable evidence that Automated Essay Scoring engines grossly and consistently over-privilege essay length in computing student writing scores. The state-of-the-art referred to in the title of the article is, in reality, simply counting words. As I have argued elsewhere (Perelman, 2012), it is this over-reliance on length that creates the apparent similarities in scores, but only for timed-impromptu writing, a genre that does not exist outside of the standardized writing test. As displayed in Fig. 1, the AES machines of the seven of nine vendors in the study that allowed their data to be released anonymously consistently overweigh word count.

The data in this figure and in both tables are reported either as correlations (the Pearson *r* product-moment correlation coefficient) or the square of the correlation, the shared variance, which is expressed as a percentage. Shared variance can be best explained as the percentage of common

variation between two variables and best represented as a Venn diagram (two overlapping circles). The shared variance is the percentage of one variable that is accounted for by the other; the overlap between the two circles.

The size and consistency of the machines' higher shared variance with word count is displayed in Table 1. In Essay Set #1, for example, the number of words determined an average of 73% of the variation in machine scores, ranging from 61.9 to 85.0%, while the two human readers had a shared variance with word count of 50.7 and 53.9% respectively. The gap between the weight given word count by machines and that given by human readers is consistent. Indeed, when comparing the shared variance between machine and human reader scores to word count, (see Table 1) there is only one instance among the 126 cases (9 scores × 7 vendors × 2 readers), in which the shared variance between a reader's score and word count is greater than that of any machine, and in that case, the difference is only 0.2%.

It is well-known that in all timed impromptu essay tests writing length comprises a significant portion of the shared variance of the scores of human readers – even the College Board concedes that a significant portion of the score on the SAT writing section essay is attributable solely to the number of words (Beckman, 2010; Kobrin, Deng, & Shaw, 2007, 2011; Winerip, 2012). Simply by overvaluing the number of words in an essay, AES machines can achieve correlations and shared variances that can, on first appearance and in some circumstances, match those achieved between two human readers.

In basing essay scores on length, machine scoring confuses association with causation. The machine algorithms – often 'black boxes' because many vendors do not disclose details of their scoring models – make the fallacious assumption that because strong student essays are usually long, long student essays are most likely strong. It is the same error in reasoning as "Many smart university professors wear tweed jackets. If I wear a tweed jacket, I will be a smart university professor." Word count is not the only variable in the regression equations that drive the calculations of AES machine but it is by far the one given most weight. Another significant factor in ETS 's E-rater regression equation, for example, is lexical complexity, which is calculated using two variables, the median number of character per word in an essay, and the use of infrequently used words measured by their frequency in popular publications (Quinlin et al., 2009).

Data refuting the principal claim in the article that machines can "consistently replicate" human scores occurs not only the raw data, but in the article itself. Examination of the tables in the article (Shermis, 2014) demonstrates that, even using the flawed methodology of the study, human readers outperformed all of the machines for some of the essay sets, particularly Essay Sets 2A and 2B, which represent scores on one of the three sets of essays that were more than a paragraph in length.

Moreover, although Shermis states that the machine scores are extremely similar to those of human scores, the analyses in this study, as I have noted in commenting on earlier versions (Perelman, 2013), compare machine scores to a construct "Resolved Score," which in half of the eight essay sets, #3, #4, #5, and #6, differs significantly from many of the readers' scores. In these four essay sets, half of the total study, if the two readers differed by one-point, the resolved score was the higher of the two, giving the machines a substantial advantage over human readers. Because the scores are discrete integer values while the essays are on a continuum (some "3's, for example, may be better than other 3's but judged to be below 4's), the machines can be programed to always round up essays that appear to be between two numbers. Human readers, on the other hand, are not scoring to match the higher of two possible scores.

Now that most of the raw test data for seven of the nine vendors has been made available (see link given above), it is possible to see how the machine scores correlate to the actual human scores, not to artificial constructs such as resolved scores. In Table 2, I display the correlations (1) between the two readers; (2) between each Vendor machine scores and the average of two readers' scores; and (3) between word count and the average of the two readers' scores. In Essay Set #2, a long persuasive essay in which readers gave less importance to word count than in other essay sets, the correlation between the two human readers was much greater than that of any of the machine's correlations. With Essay Set # 2A, the human readers shared variance to word count was only 44.1% compared to a range of 54.8–69.7% for the machines (Table 1). As shown in Table 2, the two human readers' scores for Essay score 2A, which is a holistic score for argument, organization, and development, correlate at 0.80, which when squared produces a shared variance of 63.4%. The machines' correlations with the human

Table 1
Shared variance of score to word count by vendors and human readers.

Essay Set	Vendor1 (%)	Vendor2 (%)	Vendor3 (%)	Vendor4 (%)	Vendor5 (%)	Vendor6 (%)	Vendor7 (%)	Human reader 1 (%)	Human reader 2 (%)	Average shared variance human readers (%)
1	79.9	78.9	75.6	65.8	65.2	85.0	61.9	53.9	50.7	52.3
2A	69.7	65.1	61.0	65.7	55.3	54.8	57.3	45.7	42.6	44.1
2B	48.1	39.4	39.3	55.7	37.6	37.6	37.6	24.6	24.8	24.7
3	78.8	75.8	75.1	74.1	71.4	78.4	77.2	53.1	49.9	51.5
4	69.9	69.9	74.1	70.3	67.6	73.8	72.7	55.0	55.4	55.2
5	83.8	75.9	83.2	73.6	70.5	84.0	83.6	62.6	61.6	62.1
6	57.1	53.0	57.4	55.4	42.4	60.3	54.9	38.2	42.6	40.4
7	65.0	59.7	60.8	61.2	56.8	74.7	53.1	35.6	32.5	34.0
8	48.0	41.1	48.0	40.6	36.7	39.0	29.8	19.6	18.8	19.2
Average	66.7	62.1	63.8	62.5	56.0	65.2	58.7	44.9	43.8	44.3

Source: Calculations derived from data obtained at Automated Student Assessment Prize (2013).

Table 2
Correlations–Pearson product moment correlation *r* between: (1) the two readers; (2) each Vendor machine score and the average of two readers' scores; and (3) word count and the average of the two readers' scores.

Essay Set	Readers	Vendor 1	Vendor 2	Vendor 3	Vendor 4	Vendor 5	Vendor 6	Vendor 7	Word Count
1	0.72	0.76	0.76	0.74	0.73	0.71	0.74	0.66	0.72
2A	0.80	0.71	0.72	0.69	0.70	0.71	0.68	0.68	0.66
2B	0.76	0.70	0.69	0.67	0.63	0.67	0.66	0.64	0.50
3	0.77	0.71	0.69	0.70	0.70	0.71	0.69	0.66	0.71
4	0.85	0.79	0.78	0.77	0.78	0.74	0.73	0.73	0.74
5	0.75	0.80	0.78	0.78	0.78	0.76	0.78	0.76	0.79
6	0.74	0.78	0.76	0.73	0.74	0.74	0.73	0.62	0.63
7	0.73	0.79	0.76	0.77	0.73	0.75	0.67	0.70	0.58
8	0.61	0.79	0.76	0.66	0.60	0.63	0.65	0.56	0.43

Source: Calculations derived from data obtained at Automated Student Assessment Prize (2013).

readers' range from 0.68 to 0.72. The highest scoring machine correlated with the average of the two readers at 0.72, which when squared produces a shared variance of 52.3% or 11 percentage points lower than that between the two readers. The two human readers' scores for Essay score 2B, which is a score for grammar, usage, and punctuation correlate at 0.76, while the machine's correlations with the human readers' range from 0.63 to 0.70.

In the cases in Shermis' study in which the machine scores closely match the human scores, the correspondence is attributable to one of several factors, the most important being to what extent the readers' scores correlated with word count. The two other salient factors were the nature of the writing task and the amount of noise in the computation of the writing score. As we have seen in Table 1 and Fig. 1, in Essay Set #1 the shared variance between the scores of the two human readers and word count is relatively large, 52.3%, while the shared variances between machine scores and word count were even higher, ranging from 61.9 to 85.0%, producing scores that closely matched the shared variance of the average of the two human readers. This relatively high correlation between readers and word count makes it relatively easy for AES machines to match the score of human readers simply by overvaluing word count. In this case, one successful vendor based 85% of its prediction of score on the single variable of length.

For Essay Sets #3 and #4, the nature of the writing task becomes a more important determinant than word count. Both essay prompts asked for literary analysis, which is more complex than just summarization and, consequently, the human readers outperformed all of the machines. The prompts require the writer to make inferences and interpretations of the text, cognitive tasks that human readers evaluate much better than machines, knowledge transforming rather than knowledge telling (Bereiter & Scardamalia, 1987; Deane, 2013) Although the shared variance between word count and readers' scores was substantial for both essay sets (>50%), the shared variances between the human readers were 59.8% and 71.5% respectively, substantially higher than the shared variance of any of the machines to the average score of the two human readers. On the other hand, the majority of machines outperformed human readers in Essay Sets # 5 and #6 (Table 2). As I have noted elsewhere (Perelman, 2013), the explanation is that both essays were not writing exercises at all, but reading summaries for which a list of key terms were provided to readers. Mitros, Paruchuri, Rogosic, and Huang (2013) have recently demonstrated that AES does have significant potential for the evaluation of short content and knowledge based freeform responses. Consequently, it is not surprising that some of the seven vendors have a higher correlation with the average score of the two human graders than the human graders have with each other.

For Essay Sets #7 and #8, most of the machines appeared to have greater shared variances with the average of the two reader human scores than the two readers had with each other. However, the scales used for these two essay sets were complex and subject to such a large amount of statistical noise that the all the scores can be considered to have a large random component. Essay Set # 7 consists of three analytic 0–3 scales, with the score of one of the scales being doubled to produce a 0–24 composite scale of the two readers' scores. Essay Set # 8 consists of four analytic 1–6 scales, with the score of one of the scales being doubled to produce a 10–60 composite scale of the two readers' scores. Such

composite scores are imprecise because any unreliability in each component is compounded with the addition of other components to the score. This is why the two major organizations that use analytical trait scoring of K-12 essays, Education Northwest and the National Writing Project, both employ an independent holistic score in addition to the trait scores (Swain & LeMahieu, 2012).

Another indication that the apparent accuracy of the machine scores is an artificial construct is that, with the exception of Essay Sets # 5 and # 6, human readers displayed a consistently higher level of exact agreement with each other than the vendor machine scores did with the human scores. (Shermis, 2014, p. 67, Table 7) Moreover, artificial resolved scores do not distort the metric of exact agreement because if the two readers agree, the resolved score is identical to both their scores.

There are also some other issues regarding the study that require correction or clarification:

- The conclusion to the article states that the results of the top two or three vendors were "consistently good and occasionally exceeded human rating performance." (Shermis, 2014, p. 75) However, there are no metrics given of what constituted "good" or "exceeded" and statistical tests were not performed.
- This previous issue connects to the claim that "the demonstration reported in this paper is moderated by ASAP, which acted as an independent entity with no ties or obligations to any of the developers or purveyors of machine scoring systems for essays." (Shermis, p. 54) ASAP, however, clearly did have contractual obligations with the vendors. It was not disclosed until almost a year after the initial announcement of the study that there were no statistical analyses included in the study because several of the vendors had forbidden it (Rivard, 2013).
- Describing the scoring procedures for Essay Sets # 5 and #6, Shermis states that there were a few instances in which "the state did not appear to follow its own rules in resolving the score" (p. 62) and later "were in conflict with documented adjudication procedures." (p. 74) Examination of both the test and training data sets, however, revealed that these occurrences were limited only to instances in which the two readers' scores differed by more than one point. Having a supervisor resolve such "splits" is a common and standard practice in essay assessment, (White, 1994) and almost certainly was the procedure for the two states in question. This same practice for resolving splits was consistently employed in Essay Sets # 3 and # 4. The only time the third reader reviewed a paper was when the two readers' scores differed by more than one point.
- Essay Set #7 consists of narrative essays not expository essays as reported by Shermis. The prompt for these essays is completely unambiguous. "Do only one of the following: write a story about a time when you were patient OR write a story about a time when someone you know was patient OR write a story in your own way about patience." The rubrics are also completely focused on the conventions of narrative writing (Kaggle, 2012).
- There are contradictory remarks concerning how representative these essay sets are of high stakes writing tests nationally. Section 2.1.1 of the Shermis article contains the sentence "The sample is composed of essays from volunteer states and therefore cannot be assumed to be a truly representative sample of state practice." (Shermis, 2014, p. 56) In the following section, however, there is the statement, "While there may be some debate as to whether writing samples as short as 93 words constitute an essay, these sample sizes reflect what many states are defining as essays." (Shermis, 2014, pp. 57–58) Since the five essay sets with a mean word count of less than 200 words (out of eight essay sets total) appear to come from three states, it must be asked whether this length reflects the practice of *many* states?

At the end of his article, Shermis presents a bulleted list of limitations and constraints on the study, including the failure to account for key characteristics that can affect student performance, the absence of any articulated construct for writing, and the possibility that machine scoring will change student and teacher behavior to help students game the system, what Shermis calls "signaling effects" (p. 74). Yet despite all these hedges about the limitations of his study, in the abstract Shermis makes the contradictory conclusion that "With additional validity studies, it appears that automated essay scoring holds the potential to play a viable role in high-stakes writing assessments" (p.55).

The article also claims that the performance of the top two or three vendors "was consistently good and occasionally exceeded human performance." As displayed in Table 2, when comparing vendor

machine scores to the average of human reader scores not resolved scores, the only one of the seven vendors who allowed their data to be released and consistently was among the top vendor scores is Vendor # 1, who, as shown in Table 1, also had the highest average shared variance of scores to word count, 66.7%. Moreover, because the vendors in the released data are anonymous and two of the original nine vendors chose not to allow their data to be released, we cannot even be sure that the vendors Professor Shermis referred to have allowed their data to be released and are included in this present analysis.

This confusion and ambiguity caused by a lack of transparency are central to the problems with this study. Given that the US Department of Education's Race-to-the-Top program makes student performance on these tests a major factor in school funding and the tenure, raises, and promotion of teachers, how these tests are scored will have a profound effect on American K-12 education. The machines' huge bias toward word count may encourage teachers to emphasize bloated and vapid prose. They may focus instruction on daily on-demand writing exercises to increase student output and fluency at the expense of critical thinking and frequent and extensive revision of writing. Even having the machines as second readers will produce an immense and negative bias in the scoring. As Shermis' own data demonstrates, with a normal distribution and the traditional six-point scale, adjacency, which is all that is usually necessary for a valid second read, is statistically highly probable in most cases. But the difference between an adjacent high or an adjacent low score can have a significant effect on a student's score, and compounded by a hundred or a thousand students, the machines' bias can have a profound effect on teachers and schools. Moreover, even though machines have been used as second readers on the Graduate Record Examination and Graduate Management Admissions Test, there are no published data on their effectiveness in correcting human readers.

State departments of education, state consortia, and state legislatures should carefully and independently examine the efficacy of Automated Essay Scoring in their particular contexts and for their own educational purposes. Any vendor that will not allow serious and thorough independent examinations of their AES engines should be immediately disqualified from further consideration. To do less could severely damage US K-12 education by equating assessment of essential cognitive skills with counting and writing ability with the overproduction of pointless verbiage.

References

Automated Student Assessment Prize. (2013, December). *Request ASAP phase one data*. Retrieved from ASAP: http://www.scoreright.org/asap.aspx?content=Request_ASAP_Phase_One_Data

Beckman, M. (2010, November 5). Quality vs. quantity: How to score higher on the SAT essay component: Has teen unlocked secret to a better SAT score. *Good Morning America,*. Retrieved from http://abcnews.go.com/GMA/ConsumerNews/teen-student-finds-longer-sat-essay-equals-score/story?id=12061494

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Hillsdale, NJ: Lawrence Erlbaum.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing, 18*(1), 7–24.

Higgins, J. (2014, April 24). Loss of no child left behind waiver means schools will be labeled 'failing'. *Seattle Times*, A1.

Kaggle. (2012). *Data – The Hewlett foundation automated essay scoring*. Retrieved from www.kaggle.com/c/asap-aes/data

Kobrin, J. L., Deng, H., & Shaw, E. J. (2007). Does quantity equal quality: The relationship between length of response and scores on the SAT essay. *Journal of Applied Testing Technology, 8*(1), 1–15.

Kobrin, J. L., Deng, H., & Shaw, E. J. (2011). The association between SAT prompt characteristics, response features, and essay scores. *Assessing Writing, 16*(3), 154–169.

Mitros, P. F., Paruchuri, V., Rogosic, J., & Huang, D. (2013, June 16–19). An integrated framework for the grading of freeform responses. In *Proceedings of the sixth international conference of MIT's Learning International Networks Consortium*. Retrieved from http://linc.mit.edu/linc2013/proceedings.html

Perelman, L. (2012). Length, score, time, and construct validity in holistically graded writing assessments: The case against automated essay scoring (AES). In C. Bazerman, C. Dean, K. Lunsford, S. Null, P. Rogers, & A. Stansell, et al. (Eds.), *New directions in international writing research* (pp. 121–132). Anderson, SC: Parlor Press.

Perelman, L. (2013). Critique of Mark D. Shermis & Ben Hamner, "Contrasting state-of-the-art automated scoring of essays: Analysis". *Journal of Writing Assessment, 6*(1), http://journalofwritingassessment.org/article.php?article=69.

Quinlin, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct coverage of the e-rater scoring engine (ETS RR-09-01)*. Princeton, NJ: ETS.

Rivard, R. (2013, March 15). *Professors at odds on machine-graded essays*. Retrieved from Inside Higher Ed: www.insidehighed.com/news/2013/03/15/professors-odds-machine-graded-essays

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition results and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76.

Shermis, M. D., & Hamner, B. (2012, April). *Contrasting state-of-the-art automated scoring of essays: Analysis*. Retrieved from ASAP: http://www.scoreright.org/NCME_2012_Paper3_29_12.pdf

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 213–246). New York: Routledge.
Swain, S. S., & LeMahieu, P. (2012). Assessment in a culture of inquiry: The story of the national writing project's analytic writing continuum. In N. Elliot, & L. Perelman (Eds.), *Writing assessment in the 21st century: Essays in honor of Edward M. White.* (pp. 45–68). New York: Hampton Press.
Ujifusa, A. (2014, May 7). State Political rifts sap support for common-core. *Education Week*, *33*(30), pp. 27, 30.
White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance* [E. A.-B. Edward M. White. Trans.].(2nd ed., pp. 33). San Francisco: Jossey-Bass Publishers.
Winerip, M. (2012, April 22). Facing a Robo-Grader? Just keep obfuscating mellifluously. *The New York Times*,. Retrieved from http://www.nytimes.com/2012/04/23/education/robo-readers-used-to-grade-test-essays.html

Les Perelman recently retired as Director of Writing Across the Curriculum in Comparative Media Studies/writing at the Massachusetts Institute of Technology, where he has also served as an Associate Dean in the Office of the Dean of Undergraduate Education. He is currently a research affiliate at MIT.