Trustworthy Automated Essay Scoring without Explicit Construct Validity

Patti West-Smith and **Stephanie Butler** and **Elijah Mayfield**
Turnitin, 2020 Smallman St, Pittsburgh PA 15222
{pwest-smith, sbutler, elijah}@turnitin.com

Abstract

Automated essay scoring (AES) is a broadly used application of machine learning, with a long history of real-world use that impacts high-stakes decision-making for students. However, defensibility arguments in this space have typically been rooted in hand-crafted features and psychometrics research, which are a poor fit for recent advances in AI research and more formative classroom use of the technology. This paper proposes a framework for evaluating automated essay scoring models trained with more modern algorithms, used in a classroom setting; that framework is then applied to evaluate an existing product, *Turnitin Revision Assistant*.

Introduction

Each year, millions of essays are scored automatically with models trained by machine learning, on exams like the GRE and GMAT. Historically, this industry has relied on low-dimensionality models, often using fewer than 100 features in total, constructed by researchers with psychometrics expertise. These features often represent high-level characteristics of writing like coherence or lexical sophistication[1]. This approach to model design is favored for its defensibility of the underlying model. Alignment of specific features enables “construct validity,” or rigorously defined, quantifiable alignment of model features to student behaviors that represent learning.

In modern machine learning, establishing construct validity is challenging. Machine learning researchers and practitioners are reaching a consensus that the value of a model lies in the fidelity and quantity of training data, eclipsing the value of feature engineering or hand-tuned model parameters. Fewer features are hand-crafted; in some cases, as with autoencoders, feature spaces representing text may be derived from fully unsupervised corpora (Socher et al. 2011). Following a feature-engineering approach to construct validity is not possible when using such a learned representation. Letting go of this tightly-coupled relationship between validity and representation may result in public lack of trust in automated assessment, particularly in high-stakes circumstances where skepticism is already commonplace (Markoff 2013). This slows progress in the AES field; only minimal contributions from recent research are deployed in todays automated essay scoring systems.

Today the high-stakes, high-volume use case of AES is less dominant than it once was. While high-stakes automated scoring is still widespread, most recent work has focused on applications to classrooms and student learning (Wilson and Czik 2016). Given these shifts, prior approaches to defending AES validity are becoming less informative for practitioners evaluating technology for classrooms. This paper details an alternate approach to building trust in machine learning models trained for classroom contexts. A three-pronged approach to evaluating algorithms is detailed:

- Content breadth and curriculum alignment of the product.
- Collection processes for valid, realistic training corpora.
- Scoring processes that annotate training data reliably.

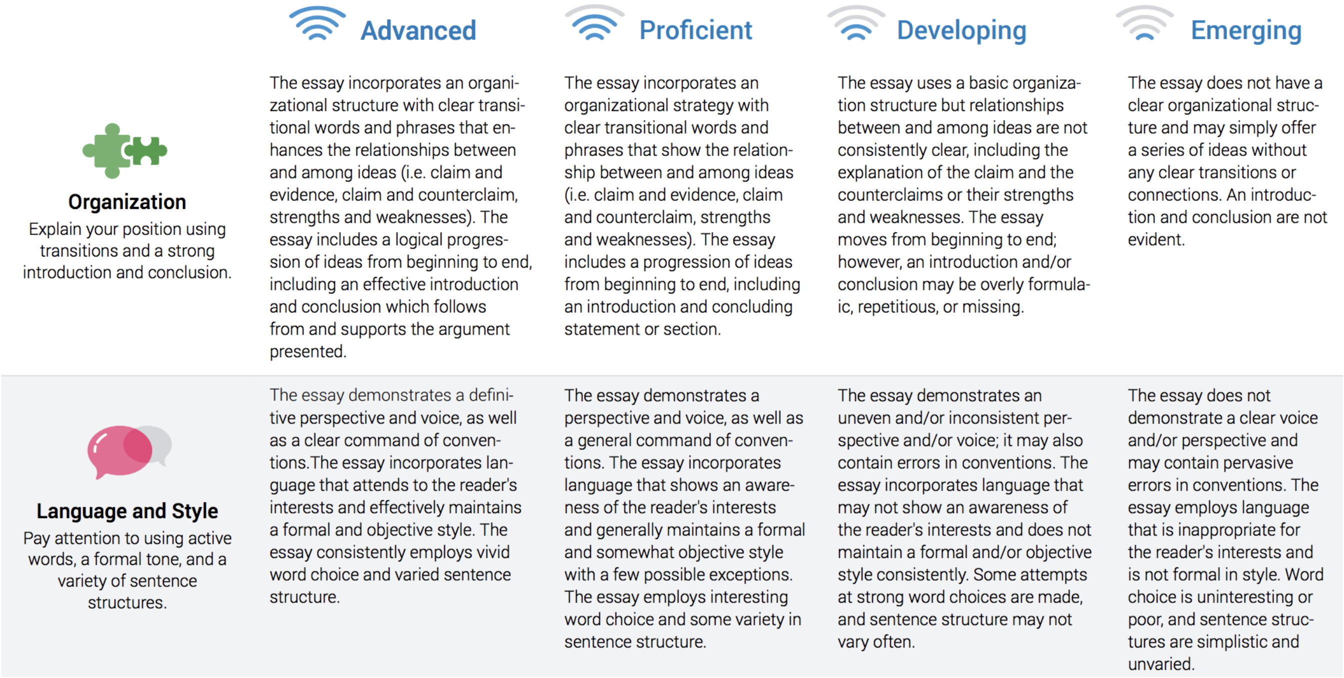
We begin with a description of the problem space, then lay out desiderata for each of the three thrusts above. The paper ends with description of one recent AES product, how these heuristics were used to inform development and deployment, and how model performance was impacted.

Problem Description

Automated essay scoring attempts to algorithmically imitate the judgment of educators evaluating the quality of student writing. Student essays are scored either on a single holistic scale, or analytically following a rubric that breaks out subscores based on “traits” (as in Figure 1). These scores are almost always integer-valued, and typically have fewer than 10 possible score points, though scales with as many as 60 points exist (Shermis 2014). In most contexts, students respond to “prompts,” a specific writing activity with predefined content, and only receive feedback on valid attempts to respond to the prompt.

An overwhelming body of evidence has shown that emulating expert scoring of essays with automated models is at least as reliable as hand scoring, or slightly better (Shermis and Burstein 2013). However, skepticism of the field remains, primarily based on the gap between “reliability” of a system - whether scores can be reproduced - and “validity”

[1] For more on the industry, see (Shermis and Burstein 2013); for details on how this is applied in practice, see (Attali and Burstein 2004).



	Advanced	Proficient	Developing	Emerging
Organization Explain your position using transitions and a strong introduction and conclusion.	The essay incorporates an organizational structure with clear transitional words and phrases that enhances the relationships between and among ideas (i.e. claim and evidence, claim and counterclaim, strengths and weaknesses). The essay includes a logical progression of ideas from beginning to end, including an effective introduction and conclusion which follows from and supports the argument presented.	The essay incorporates an organizational strategy with clear transitional words and phrases that show the relationship between and among ideas (i.e. claim and evidence, claim and counterclaim, strengths and weaknesses). The essay includes a progression of ideas from beginning to end, including an introduction and concluding statement or section.	The essay uses a basic organization structure but relationships between and among ideas are not consistently clear, including the explanation of the claim and the counterclaims or their strengths and weaknesses. The essay moves from beginning to end; however, an introduction and/or conclusion may be overly formulaic, repetitious, or missing.	The essay does not have a clear organizational structure and may simply offer a series of ideas without any clear transitions or connections. An introduction and conclusion are not evident.
Language and Style Pay attention to using active words, a formal tone, and a variety of sentence structures.	The essay demonstrates a definitive perspective and voice, as well as a clear command of conventions.The essay incorporates language that attends to the reader's interests and effectively maintains a formal and objective style. The essay consistently employs vivid word choice and varied sentence structure.	The essay demonstrates a perspective and voice, as well as a general command of conventions. The essay incorporates language that shows an awareness of the reader's interests and generally maintains a formal and somewhat objective style with a few possible exceptions. The essay employs interesting word choice and some variety in sentence structure.	The essay demonstrates an uneven and/or inconsistent perspective and/or voice; it may also contain errors in conventions. The essay incorporates language that may not show an awareness of the reader's interests and does not maintain a formal and/or objective style consistently. Some attempts at strong word choices are made, and sentence structure may not vary often.	The essay does not demonstrate a clear voice and/or perspective and may contain pervasive errors in conventions. The essay employs language that is inappropriate for the reader's interests and is not formal in style. Word choice is uninteresting or poor, and sentence structures are simplistic and unvaried.

Figure 1: Two traits from a rubric designed for use with an AES system.

of a system - whether the predicted scores are representative of student ability rather than superficial correlates. AES researchers have sometimes claimed that reliable reproduction of scoring by expert judges is itself sufficient evidence of validity, *"recogniz[ing] the primacy of human judges as the most important criterion to emulate"* (Keith 2004). These defenses have largely been dismissed by writing assessment scholars as inadequate (Perelman 2014).

Other defensibility arguments have focused on the expert judgment in feature engineering of AES models. In 2004, a defense of then-leading automated scoring model, ETS e-Rater, argued that its 12 features *"reflect essential characteristics in essay writing and are aligned with human scoring criteria [...] Validity here refers to the degree to which the system actually does what is intended, in this case, measuring the quality of writing."* (Burstein, Chodorow, and Leacock 2004). This has been taken more seriously and led to use of these systems in high-profile standardized exams.

These arguments have been insufficient for convincing teachers to use AES in classrooms. Teachers using tools derived from this research have viewed the psychometric models as "fallible" (Grimes and Warschauer 2010) and stated that automated scoring must be paired with actionable next steps for writers (Riedel et al. 2006). Based on this feedback, writing instruction tools based on AES has been studied closely for use in "formative" learning applications, rather than "summative" scoring-only settings. The major differentiator is the presence of automated feedback and the chance for students to revise their work based on that feedback in real-time. Automated feedback in this category has been percevied by students as informative, valuable, and enjoyable (Roscoe et al. 2014) and which provided more efficient learning gains than practice alone (Crossley et al. 2013). To date, these values have not been well-described or evaluated by psychometric validity arguments. To date, no systematic framework for evaluating a model's fit for learning purposes has been adopted in either academic or industry applications.

In response to this gap, the following three sections describe defensible practices for training AES models for a classroom setting. Similar blueprints for evaluation of deployed models have been described more broadly for machine learning systems, from engineering (Sculley et al. 2015) to annotation (Mason and Suri 2012)[2], but not in the education domain. The first sections describes "Curriculum Validity," the selection of content for production use of an AES system, based on a collaboration with practicing educators. The second describes "Data Validity," authentic collection of student samples for training sets, relying on partnership with teachers (and their students). The final section describes "Annotation Validity," a process for highly reliable scoring, based on close collaboration on defining the labels for training sets. This framework evaluates the quality of an AES system based on the process that led to its curriculum, its essays, and their scores, rather than on expert feature engineering or interpretability of model weights.

Curriculum Validity

It is well-established that there are gaps between instructional effectiveness research, the authoring of curriculum

[2] The cited framework specifically focuses on Amazon Mechanical Turk, but has been applied more broadly.

materials, and the application of those materials by practicing educators (Ball and Cohen 1996). AES products, especially those designed for summative purposes, are particularly vulnerable to this gap. Existing models, in general, have not adapted their content as education standards have shifted. For instance, Latent Semantic Analysis is well-targeted to summarization tasks; this technical approach predates the current Common Core Standards by more than a decade, yet is still a primary component of modern AES products (Foltz, Hidalgo, and Van Moere 2014).

The goal of an AES prompt library should be to allow teachers to use formative writing feedback in varied settings throughout an academic year, giving students feedback that supports progress over time and across genres. School districts bring critical expertise for choosing the materials necessary to achieve this goal. Content for writing assignments is more applicable to classrooms when collaborating with practitioners, including choice of reading materials and alignments to grade level, content area, genre, and accountability standards. This content must then be evaluated based on the technical constraints. This section recommends practices that lead to AES prompts that meet these goals.

1. Authentic sourcing from educators. As preliminary steps, school districts and practicing teachers should determine relevant content areas for use in an AES prompt library, including subject and source materials (if any). The intended purpose of the content should be recorded - for instance, benchmark essays for a start-of-year assessment fulfill a different purpose than a low-stakes practice essay as part of a multi-day instructional activity. At this initial review stage, prompts should be authored and some small number of sample essays - typically fewer than ten for each prompt - are gathered for interdisciplinary review in the next steps below. District partners provide scores for these sample essays if available, either by trait or holistic.

2. Machine learning capacity for evaluation. Machine learning practitioners are responsible for assessing whether a prompt is appropriate and capable of assessing a prompt. Warning signs of incompatibility can include an overly broad topic, which can result in overly varied and ambiguous training sets of sample essays; constrained, non-prose writing forms, including most poetry; and document length and format, where highly structured documents and multi-modal content may break expectations of machine learning feature extraction. Notably, the inclusion of poetry in *source materials* for prompts is not in itself a red flag, as student analysis of that content is typically still within the capabilities of AES models.

3. Library diversity. Expansion of prompts in an AES library should be evaluated in the broader context of existing content. Recreating new, overly similar content can slow teacher lesson preparation with ambiguous materials. Providing a wide range of options while maintaining organized, clear boundaries between prompts with varied content and goals, by contrast, benefits teachers. The prompt and sources must also be reviewed for clarity among diverse student populations; region-specific language or vocabulary, for instance, has the potential to widen pre-existing gaps in achievement.

4. Education standards and accountability context. For use *in situ*, support for teachers subject to accountability measures must also be considered. Practicing teachers are held to strict expectations, such as the Common Core or equivalent state-specific standards. Support can come from materials like *crosswalks*, a document that allows line-by-line comparison between a source rubric and a comparison set of standards (see Figure 2). Crosswalks are convenient in that they allow a one-to-many relationship, with a single well-designed rubric aligning to multiple state standards and reducing needless replication of expert-authored materials. Essay prompts may be categorized in theses systems, often by grade band and genre - for instance, middle school argument, or high school text-based analysis. Prompts can also be grouped at a higher-level abstraction (for example, aligned to "essential questions" or "scope and sequence" documents that are common in textbooks used in schools). Content from well-known 'canon' texts, such as *Hamlet* or *To Kill A Mockingbird*, may require less support than obscure or original texts.

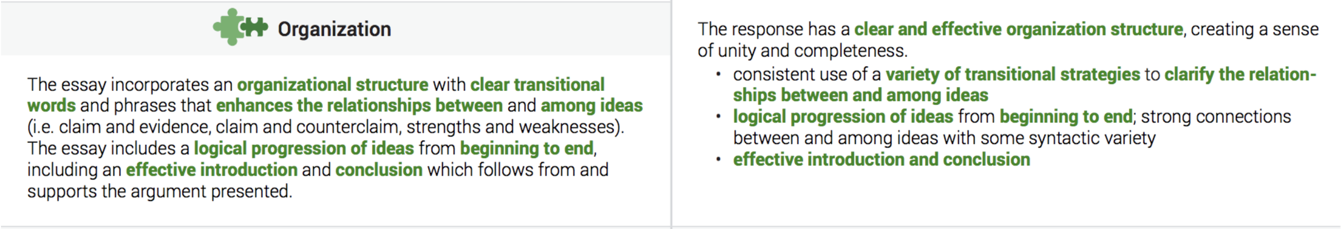
5. Relevance and recency of materials. Source materials should be relevant to students' daily lives and experiences. However, the most relevant and timely source materials are often under copyright and AES engines must determine copyright permission status for prompts and source materials if they are to be included in a curriculum and then redistributed. Materials that are out of copyright should be explained in their contemporary context for students without that pre-existing background understanding.

6. Disciplinary literacy. Typically, writing assignments along with reading are thought of as part of an English Language Arts curriculum in American schools. However, this is not the only place where AES has applications. Disciplinary literacy is an appropriate use of technology in social studies, physical sciences, or other fields, so long as all other constraints are still met. This widens the scope of the technology beyond what is typically discussed in the literature. In fact, disciplinary text-based responses are often more well-suited to fact-based analysis than more argument-oriented texts, and AES has been shown to reliably score these questions based on the student's grasp of higher-level generalizations (Nehm, Ha, and Mayfield 2012).

Data Validity

AES models are trained through supervised machine learning. This requires a collection of student responses to build a corpus for each prompt. These responses are collected well in advance of use of an AES model either in formative or summative settings, but can be collected in ways that produce poor-quality datasets, non-representative subsets of student writing, or unmotivated student responses.

Student essays should represent a broad spectrum of authentic student attempts at responding to a prompt, demonstrating their true writing ability, across a wide array of students. Inappropriate collection of data results in inaccurate evaluation of new submissions by an AES model. For example, in the commonly-used gold standard dataset ASAP (Shermis 2014), essays were largely authored in a standardized testing setting. Students were expected to author essays



Organization

The essay incorporates an **organizational structure** with **clear transitional words** and phrases that **enhances the relationships between** and **among ideas** (i.e. claim and evidence, claim and counterclaim, strengths and weaknesses). The essay includes a **logical progression of ideas** from **beginning to end**, including an **effective introduction** and **conclusion** which follows from and supports the argument presented.

The response has a **clear and effective organization structure**, creating a sense of unity and completeness.

- consistent use of a **variety of transitional strategies** to **clarify the relationships between and among ideas**
- **logical progression of ideas** from **beginning to end**; strong connections between and among ideas with some syntactic variety
- **effective introduction and conclusion**

Figure 2: A sample crosswalk between 9-10th grade *Argument* rubric and 9-10th grade Smarter Balanced consortium standards.

in artificial, timed, closed-notes settings. This can lead to bad-faith submissions:

> *"Scientests at the* `@CAPS7` *lab in* `@LOCATION2` *said that* `@NUM1` *out of* `@NUM2` *regular computer users lost their vision within two years. One of these scientist,* `@PERSON3`*, reported, "*`@CAPS8` *more people begin to use the computer, more people seriously hurt their eyes or even lose their vision. We estimate* `@PERCENT2` *of this next generation will be legally blind before age* `@NUM3`*."*

Even in major public datasets, inauthentic student writing is rife with references to invented quotes and fabricated research[3]. The essay from which this excerpt came received a score of 11/12 in the ASAP dataset; all AES systems using this corpus are therefore trained to recognize such writing as high-quality. If a training set contains bad-faith or unmotivated essays, it adversely impacts the potential of an AES writing intervention. The frequency of such essays being included can, however, be mitigated by following established protocols. The recommendations below keep students and teachers engaged during training set collection, resulting in a wider range of student abilities.

1. Collect from diverse populations. A wide range of student writers should be present in a training set, representing most or all common responses to a prompt. Oversampling from a narrow population of similar students makes this representation less likely. This step ensures that many potential approaches are represented when responding to a prompt, rather than only the default expectations of teachers. This also helps ensure all possible scores appear in a training set, on each trait; it is difficult to create reliable models that can provide appropriate feedback if some student groups do not appear in training data.

2. Intentionally oversample tails. Some student populations are smaller by nature; in a normal distribution, receiving a score at the floor or ceiling of a trait's range is rare by definition. Fewer of those responses will appear in a uniform sample of student responses. It is often appropriate to assign collection to specified subsets of classrooms that are more likely to elicit writing at each score point. In some circumstances, when a prompt is particularly difficult, it may be appropriate to collect a small number of responses with an advanced group, or even a slightly older group of students, to build a representative sample at the tails of a distribution. The same can be true at the scoring floor, which may benefit from collection from slightly younger students.

3. Avoid student fatigue. The end goal of a collection process is to increase the breadth of a content library; this may result in students being asked to write in response to multiple prompts if a school district is a partner on a large set of prompts. However, not all training sets can come from the same group of students, especially in a relatively short period of time. When students are asked to write repeated assignments (especially without significant feedback in between), quality decreases. That slump leads to lower scores, more shortcuts by students, and a narrower range of responses. Slower collection of prompt datasets over time maintains high standards for quality.

4. Make motivations clear. How teachers view a collection process will impact the way they depict that process to students. This impacts the quality of responses. Teachers (and ultimately, students) should have the end goal for the collection clearly articulated to them. When teachers are unsure, they sometimes believe that the process is an accountability measure on *their* teaching, or the collection may be seen as a distraction from other instructional content. When that happens, their feelings bleed through to their students, who are then less motivated; in the worst case, students may use their essays as an outlet to complain about the classroom process. This is exacerbated when a group of students responds to multiple prompts in a brief window, as in the fatigue point above.

5. Avoid scrubbing the data. It is natural for data scientists to remove atypical responses or early drafts, which can be seen as noise. Sometimes, district partners want to give only their best, exemplar responses from students. In that effort to "look good," they end up not supplying a complete set of essays or all the associated data. Districts sometimes use a prompt with multiple groups of students, but only provide essays that "fit" how they view successful responses to the prompt. By doing this, students at the low end of performance are intentionally omitted from representation. In practice, all sampled essays, including those in the tails of scoring, help in training. In general, a larger set of essays gives more for a model to learn from, leading to a model that can provide feedback to a broader range of students. Additionally, writing at different stages of completion is likely to appear within the live context of an AES intervention, and

[3] Named entities are anonymized in public data releases and this excerpt, but the inaccuracy of this and other examples in the ASAP dataset has been confirmed through personal communication. For details on anonymization's impact on AES reliability, see (Shermis, Lottridge, and Mayfield 2015).

should not be totally foreign to the trained model. When logistically possible, collectors should collect early drafts of student work for scoring, to represent growth in essay quality over time. Filtering essays to remove outliers can be time-consuming and counter-productive.

6. Overcommunicate with partners. Many of the above heuristics overlap. A well-documented plan for content collection will help all parties anticipate these issues and problem solve if factors could negatively impact the quality of the set. A collection process does not necessarily need to receive all data at one time; assigning batches of essays at typical peak writing periods during the school year yields higher-quality training sets.

Annotation Validity

Typically, scoring of large corpora of collected student text is not completed either by educators who collect those essays, or by AES practitioners. Instead, it is treated as a supervised annotation task and outsourced to a third party, consisting of large numbers of moderately trained participants and a smaller number of "lead scorers" with more experience and decision-making authority. Data is transferred to such a vendor after accounting for privacy regulations and removal of personally identifiable information. It takes work to build a relationship with a scoring vendor. If either side is not open to discussion and feedback, the partnership is not likely to meet the needs of both sides and will almost certainly not support a reliable partnership. This section specializes established best practices on corpus annotation[4] to the domain of rubric-based scores on student essays; in this context, the terms annotation and score are interchangeable.

1. Establish a collaborative process. Feedback and concerns from annotators are integrated into the scoring process. A vendor should read representative subsets of essays for scoring prior to large-scale annotation, and flag potential problems and requests for clarification. For instance, alignment between prompts and rubrics should be clear to vendors. If a set of essays does not match the expectations of the rubric, it should be identified upfront. Sometimes, this may be remedied with a clear rationale from the provider of collected data; in other cases, severe problems may lead to changes to a scoring rubric itself.

2. Identify anchor papers. In order to consistently apply a rubric to essays written to a specific prompt, an anchor paper review is crucial. In this process, scoring leads identify "anchor papers" that exemplify the score points across a rubric; these anchor papers are used to train the individual annotators. This process should be two-sided between researchers and the vendor; one group should submit their proposed set to the other group for consensus-building, to pair scoring expertise with knowledge of classroom context. This process should be iterative and anchor papers typically are added or removed through discussion prior to training annotators.

3. Develop clearly articulated rubrics. Clear lines should be drawn between performance levels in traits of a scoring rubric. Traits themselves must be distinguished from one another. Cross-correlated expectations across traits harm scoring quality. For example, the use of transition words may impact the overall quality of organization, and might also help to show the relationships between the claim and the evidence used to support it, but a rubric should be designed with each aspect of writing isolated to one trait. Annotators should be trained where students get "credit" for a particular skill. Additionally, rubrics must articulate a stepwise progression upwards through score points. The language that maps out what occurs within a trait has to be developed with key criteria for students and teachers, as well as alignment to accountability standards (as discussed above).

4. Build in a common vocabulary. Rubrics that map out score points often contain subjective phrases. Modifiers like "significant" or "thorough" can be interpreted differently by individual annotators. The distinction implied by these terms should be explained during training of annotators, and when reused, should have consistent meaning across traits. Descriptive language, particularly adverbs, should not vary in meaning across rubric traits or score points. Using anchor papers is a useful step in the process of defining these key modifier words, as they can be tied to authentic examples.

5. Use specific examples from student work. Essays collected authentically, following the Data Validity steps above, should be used in the collaboration on scoring best practices. Abstract ideas represented in a rubric should be rooted in real student writing to make them concrete. For example, "an objective tone" in a middle school essay collection is difficult to describe by adults, and may be easier to describe through examples of middle school text. Commentary by either researchers or vendors may be attached as qualitative explanations on student text during training, for rationale and clarity, but the concrete representation of concepts is more vital. Because there is no single correct way to construct an essay, multiple examples are often clarifying.

6. Avoid latent language ideologies. Student writing is produced in response to prompts that outline the language expectations for the assignment. In turn, annotators should score student responses based only on those explicit requirements. Annotators' personal preference or cultural familiarities may alter their holistic perceptions of writing quality. This can be expressed through subtle style biases, such as through dialect markers or grammaticality, or through hidden structural requirements like minimum word counts. Such subtle biases can disproportionately impact protected classes and students of color (Godley et al. 2006). This can undermine the validity of scoring, and it is therefore important to limit training of annotators to focus on the identified, specific writing requirements that were given to students during data collection.

7. Systematically evaluate scoring output. Consistently evaluate the scores given to each dataset before accepting scoring as complete. Design this evaluation system to capture the most common problems with calibration or misalignment of the scorers, and also the most common dataset flaws introduced by the data collection process. Scorer and data problems will likely be confounded, so an expert may need to determine the proper course of action in the case of poor results. The most commonly flagged error patterns

[4] See for example (Hovy and Lavid 2010).

include rare representation at the ends of a scale, extreme over-representation of a single score point, or poor agreement of individual annotators who are "out of sync" with the rest of a group. Unusually strong correlation with essay length, or cross-correlation between essay traits, is also a sign of rushed scoring.

8. Share expectations around hiring and onboarding. Any vendor that works with essay scoring will have a standing process for selecting scorers, training, and calibration. This process is typically separate from, and in addition to, the scoring process for an individual dataset. For any organization working with a scoring vendor, it is essential that the organization has transparency into those established procedures. Beyond that, though, it is also important to make sure that both sides of the working relationship have a shared understanding of what processes are put in place to make sure that scoring leads are effectively chosen from a broader group of scorers, as they are the ones who must effectively train and disseminate critical information to the actual scorers. For valid scoring of training sets, there must be trust that scorers have experience and expertise in scoring the written work of adolescent students.

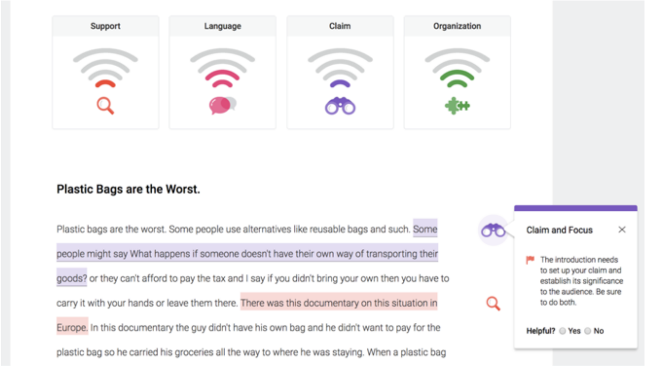
Evaluation in Practice

This remainder of this paper applies this framework to the evaluation of *Revision Assistant*, an AES intervention developed by Turnitin and primarily designed for formative classroom use and deployed at scale in American middle and high schools. *RA* emphasizes the importance of the writing process by reframing essay authorship as an on-going activity. The design of the system utilizes AES to embed an intensive revision process into student interactions with the system.

As students request automated scoring, feedback is also provided; *RA* highlights two relatively strong sentences and two relatively weak sentences (Woods et al. 2017). Instructional content appears alongside those sentences that helps students understands where they are excelling in their writing and where they should focus their revision efforts. Comments encourage students to take small, targeted steps toward iteratively improving their writing.

This design and pedagogical constraint is meant to provide students with the opportunity and the desire to engage in writing strategies around constant refinement and iteration. By creating an environment that directly connects student writing to feedback that encourages rework, it becomes clear to the student that good writing is the product of multiple drafts. The instantaneous nature of the feedback further aids students by creating an environment where revision can easily take place. Feedback cycles which could be days or weeks long are shorted to near-instantaneous feedback. This makes it significantly more motivating for students to revise and improve their work. The visual, game-like appeal of Wifi signals creates an atmosphere that encourages students to work and improve, without the feeling of finality from previous, summative AES systems.

Evaluating Curriculum Validity



Plastic Bags are the Worst.

Plastic bags are the worst. Some people use alternatives like reusable bags and such. Some people might say What happens if someone doesn't have their own way of transporting their goods? or they can't afford to pay the tax and I say if you didn't bring your own then you have to carry it with your hands or leave them there. There was this documentary on this situation in Europe. In this documentary the guy didn't have his own bag and he didn't want to pay for the plastic bag so he carried his groceries all the way to where he was staying. When a plastic bag

The introduction needs to set up your claim and establish its significance to the audience. Be sure to do both.

Figure 3: The user interface of *Revision Assistant*.

Content within *RA* is wide-ranging (see Table 1). At time of this paper's authoring, content is distributed across genres, subject areas, and grade levels, though with more comprehensive in high school grade levels. Content is weighted towards English Language Arts. A subset of prompts has been specifically identified as appropriate for summative assessment purposes, while the rest are recommended for lower-stakes use only. This constitutes an appropriately diverse library with room for improvement in the physical sciences and in the younger grades.

RA's scoring rubrics are genre- and grade-band dependent, but do not vary across prompts within those genres and grades. The rubrics are also designed for alignment with crosswalks to four different standards consortia - Smarter Balanced, PARCC, Texas Education Agency, and Florida Department of Education. Including all states which adopted the Common Core, this results in accountability crosswalks for 46 states and DC[5].

Source texts are weighted towards modern writing, with more than half of sources written in the 21st century. A wide range of identities are represented in source texts, including African American, Native American, Asian-American, and Hispanic authors, as well as texts by non-American authors. There is room for broader inclusion - fewer than 10% of texts are authored by women of color, and zero are written by nonbinary gendered authors.

Overall, *RA* provides strong curriculum validity for practitioners, including a wide-ranging library, alignment to standards in most school districts in America, and modern, diverse representation in authorship.

Evaluation Data Validity

For most prompts, collection of student work for *RA* training sets was timed across multiple months and in line with regular teaching practices.

Once tasks and rubrics were established with partners, best practices as described above were shared with district partners. For library expansion in the 2016-2017 school year, collections were timed in eight "waves" across 20 school districts, expanding the library by at least 50 prompts. Each wave consisted of between one and four participating school districts. Each wave consisted of pilot use of the

[5] Education standards in Alaska, Nebraska, Oklahoma, and Virginia are not well-aligned to content in *Revision Assistant*.

Genre	Narrative Writing	11
	Informative Essays	31
	Text-based Argument	12
	Open-ended Argument	10
	Textual Analysis	17
Subject Area	English Language Arts	78
	Social Studies / History	32
	Physical Sciences	20
Grade Level	Grade 6	14
	Grade 7	22
	Grade 8	24
	Grades 9-10	32
	Grades 11-12	28
Summative Use	Timed High-Stakes OK	42
	Low-Stakes Only	42
Source Text Date	Before 1900	19
	1901-2000	33
	Since 2000	96
Expressed Identity of Source Author	Women (of color)	41 (10)
	Men (of color)	80 (16)
	Not presented	42

Table 1: Prompt library and source text distributions in *RA*.

prompt in classroom settings, evaluation of initial essays, and then a larger collection process across more classrooms.

Because of the nature of the work, waves were staggered, sometimes over a number of months. Though standards and pacing may provide guidelines for curricular materials, not all teachers assigned work on the same day, or even the same week. Each training set was collected by between 5 and 10 teachers, and waves consisted of a minimum of 5 and a maximum of 25 distinct prompts. Typically, groups of teachers within buildings were part of each wave, rather than working with individual teachers isolated from the process.

Datasets were vetted based on minimum training set size targets. Districts did not appear in waves as the sole participating district unless at least 500 unique student essays, spread across score points, could be reliably collected for each training set in that wave. This barrier prevented single-district training set collections in most cases. Instead, multiple districts share training set responsibility for each prompt, in order to ease the burden of collection on any one district. Essays were collected either through *RA* in an interface with no automated feedback, or were collected in other word processors and shared over secure file transfer.

Evaluating Annotation Validity

At a high level, performance of the model resulting from this collection process is measured through Quadratic Weighted Kappa, or QWK, the industry-standard method of evaluating model quality (Shermis and Burstein 2013). On this metric, which typically ranges from 0 to 1, industry best practices recommend performance of at least 0.6 before use even in low-stakes settings, and an optimal target of up to 0.8 for ”near-perfect” reproduction of expert scores. Further detail on model performance can be gleaned through evaluated score distributions across a training set's true and cross-validated predicted labels, and confusion matrices that highlight frequent mismatches between scores.

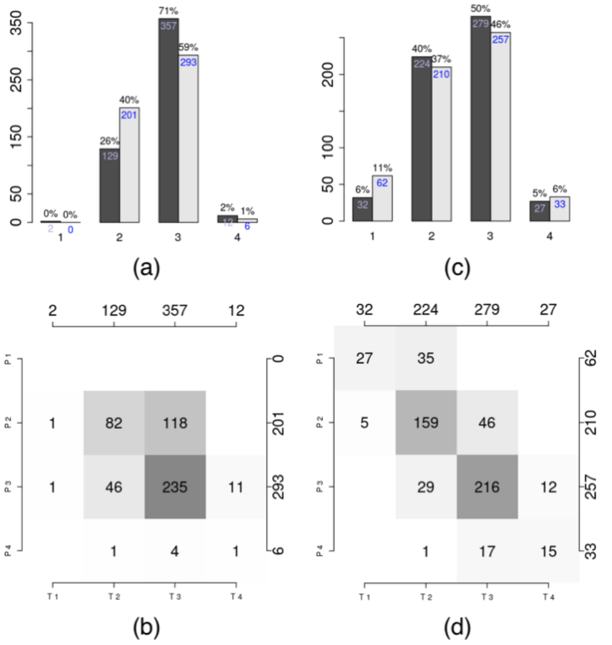


Figure 4: Analysis of a single prompt on a single trait before (a-b) and after (c-d) best practices for training set collection were put in place.

Figure 4 illustrates these visualizations in a model trained to score a 9th-10th grade essay prompt. Two sets of scored essays are shown, before and after application of the best practices above. Evaluation is completed through 10-fold cross-validation of a training set of 490 essays. Prior to implementation of best practices, the first model reaches only a QWK of 0.281, well below the industry benchmark. The final model's QWK reaches 0.749, above the threshold for high-stakes use. A more in-depth quantitative analysis is instructive and highlights the problems in hand-scoring when best practices are not followed. The distribution in the top left (a) presents counts of scores within the training set, both in ground truth (dark) and predicted (light) score sets. In the original hand scoring, 71% of all essays received the most common score of 3/4, and only 2 essays received the minimum score of 1/4. This "clumping" to the middle is common when oversight is minimal. The confusion matrix (b) highlights the challenge of automated scoring when essays are scored this way. The model learns to replicate observed scoring behavior, and ignores both the top and bottom of the scoring range altogether. Even within the two frequent score points, confusion is common; fewer than 65% of essays are scored exactly correctly, worse than would be expected from a trivial classifier that always predicted the majority class.

The right-hand column smooths out these problems somewhat. As seen in the score distribution (c), scores at the top and bottom of the scoring range now account for more than 10% of essays in the training set, enough for machine learning algorithms to identify reliable characteristics of 1/4 and 4/4 scores. The confusion matrix (d) shows that all four score points can now be reliably identified, even though the majority class still accounts for half of all essays. No errors greater than adjacent misses are made at any point during cross-validation. This pattern of improvement indicates a material improvement in scoring behavior as a result of the practices described in this paper.

Conclusion

Educators should expect AES to be held to a high standard when selecting interventions for use in classroom settings. Transparency in content selection, curriculum alignment, training set collection practices, school partnerships, and annotator hiring and training form a broad and comprehensive picture of automated essay scoring model behavior. This picture exceeds the transparency typical in the psychometric literature, which only gives sparing coverage to qualitative aspects of model training and emphasizes reliability.

Following best practices on all three categories - curriculum, data, and scoring - requires an extended partnership between school teachers, machine learning researchers, and annotators. This is a more interdisciplinary approach than statistics-driven arguments for validity, and requires more transparency than the AES community has previously been subjected to. Models trained at the end of a process that follows these best practices, however, both provide reliable scoring of student essays and support classroom instruction.

References

Attali, Y., and Burstein, J. 2004. Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series* (2).

Ball, D. L., and Cohen, D. 1996. Reform by the book: What isor might bethe role of curriculum materials in teacher learning and instructional reform? *Educational researcher* 25(9):6–14.

Burstein, J.; Chodorow, M.; and Leacock, C. 2004. Automated essay evaluation: The criterion online writing service. *AI Magazine* 25(3):27.

Crossley, S.; Varner, L.; Roscoe, R.; and McNamara, D. 2013. Using automated indices of cohesion to evaluate an intelligent tutoring system and an automated writing evaluation system. In *International Conference on Artificial Intelligence in Education*. Springer.

Foltz, P.; Hidalgo, P.; and Van Moere, A. 2014. Improving student writing through automated formative assessment: Practices and results. In *International Association for Educational Assessment Conference*.

Godley, A.; Sweetland, J.; Wheeler, R.; Minnici, A.; and Carpenter, B. 2006. Preparing teachers for dialectally diverse classrooms. *Educational Researcher* 35(8):30–37.

Grimes, D., and Warschauer, M. 2010. Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning, and Assessment* 8(6).

Hovy, E., and Lavid, J. 2010. Towards a scienceof corpus annotation: a new methodological challenge for corpus linguistics. *International journal of translation* 22(1):13–36.

Keith, T. 2004. Validity of automated essay scoring systems. In Shermis, M., and Burstein, J., eds., *Automated essay scoring: A cross-disciplinary perspective*. Lawrence Erlbaum Associates. 266–290.

Markoff, J. 2013. Essay-grading software offers professors a break. *The New York Times* A1.

Mason, W., and Suri, S. 2012. Conducting behavioral research on amazons mechanical turk. *Behavior research methods* 44(1):1–23.

Nehm, R.; Ha, M.; and Mayfield, E. 2012. Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *Journal of Science Education and Technology* 21(1):183–196.

Perelman, L. 2014. When “the state of the art” is counting words. *Assessing Writing* 21:104–111.

Riedel, E.; Dexter, S. L.; Scharber, C.; and Doering, A. 2006. Experimental evidence on the effectiveness of automated essay scoring in teacher education cases. *Journal of Educational Computing Research* 35(3):267–287.

Roscoe, R.; Allen, L.; Weston, J.; Crossley, S.; and McNamara, D. 2014. The writing pal intelligent tutoring system: Usability testing and development. *Computers and Composition* 34:39–59.

Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.-F.; and Dennison, D. 2015. Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems*.

Shermis, M., and Burstein, J. 2013. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Shermis, M.; Lottridge, S.; and Mayfield, E. 2015. The impact of anonymization for automated essay scoring. *Journal of Educational Measurement* 52(4):419–436.

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.

Socher, R.; Pennington, J.; Huang, E.; Ng, A.; and Manning, C. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of Empirical Methods in Natural Language Processing*.

Wilson, J., and Czik, A. 2016. Automated essay evaluation software in english language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education* 100:94–109.

Woods, B.; Adamson, D.; Miel, S.; and Mayfield, E. 2017. Formative essay feedback using predictive scoring models. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*.