Automated Evaluation of Writing – 50 Years and Counting

Beata Beigman Klebanov and Nitin Madnani
Educational Testing Service, Princeton, NJ, USA
{bbeigmanklebanov,nmadnani}@ets.org

Abstract

In this theme paper, we reflect on the progress of Automated Writing Evaluation (**AWE**), using Ellis Page's seminal 1966 paper to frame the presentation. We discuss some of the current frontiers in the field, and offer some thoughts on the emergent uses of this technology.

1 A Minimal Case for AWE

In a seminal paper on the imminence of automated grading of essays, Page (1966) showed that a high correlation between holistic machine and human scores is possible. He demonstrated automated scoring of 276 essays written by high school students by a system with 32 features, resulting in a multiple $R = 0.65$ between machine and average human score, after adjustment. He also provided a thoughtful discussion of his ambitions for automated scoring and of the possible objections.

Page made the case that automated evaluation of student writing is needed to take some of the evaluation load off the teachers and to provide students evaluations of their (potentially multiple) drafts with a fast turnaround. He then appealed to the then-burgeoning interest and fascination with machine learning to argue for the feasibility of such an enterprise, namely, that machines can learn how to give the right grades to essays, if trained on an expert-scored sample.

As part of the feasibility argument, Page emphasized the need to carefully define the goal so that success can be judged appropriately. The goal is not a "real" master analysis of the essay the way a human reader would do but merely an imitation that would produce a *correlated* result (using what Page called *proxes* – approximations). Page considered this goal to be both useful and achievable.

2 Report Card: Where are We Now?

2.1 Accomplishments

Page's minimal desiderata have certainly been achieved – AWE systems today can score in agreement with the average human rater, at least in some contexts.[1] For example, Pearson's Intelligent Essay Assessor™ (IEA) scores essays written for the Pearson Test of English (PTE) as well as for other contexts: "IEA was developed more than a decade ago and has been used to evaluate millions of essays, from scoring student writing at elementary, secondary and university level, to assessing military leadership skills."[2] Besides sole automated scoring as for PTE, there are additional contexts where the automated score is used in addition to a human score, such as for essays written for the Graduate Record Examination (GRE®)[3] or for the Test of English as a Foreign Language (TOEFL®).[4] Does this mean that the problem of AWE is solved? Well, not exactly.

2.2 Needs Improvement

Page did anticipate some difficulties for AWE systems. It is instructive to see where we are with those.

2.2.1 Originality

> *What about the gifted student who is offbeat and original? Won't he be overlooked by the computer?* (Page, 1966)

Page's argument is that the original student is not going to be much worse off with a com-

[1] It is not our goal to survey in detail techniques that underlie this success. See Ke and Ng (2019) for a recent review.

[2] https://pearsonpte.com/the-test/about-our-scores/how-is-the-test-scored/

[3] https://www.ets.org/gre/revised_general/scores/how/

[4] https://www.ets.org/toefl/ibt/scores/understand/

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7796–7810
July 5 - 10, 2020.

puter than with an (average) human reader, because originality is a subjective construct. Thus, once research uncovers objective and measurable aspects of "original" writing, relevant features can be added into an AWE system; finding such aspects, as well as measuring them, is still work in progress. While no current operational scoring system we are aware of is specifically looking for originality, research into aspects of writing that are often considered original is taking place. For example, using data from different tests, Beigman Klebanov and Flor (2013a) and Beigman Klebanov et al. (2018) found that the extent of metaphor use (proportion of metaphorically used words in an essay) correlates with essay quality; Littlemore et al. (2014) likewise found that more skilled writers use metaphor more often. Song et al. (2016) observed a positive correlation between use of parallelism – syntactically similar and semantically related constructors, often used for emphasis or to enhance memorability – in student essays. Some pioneering work has been done on comparing writing that is recognized as outstanding (through receiving prestigious prizes) vs writing that is "merely" good in the domain of scientific journalism (Louis and Nenkova, 2013). Once various indicators of originality can be successfully measured, additional work may be necessary to incorporate these measurements into scoring ecosystems since such indicators may only occur infrequently. One way to achieve this would be to compute a "macro" feature that aggregates multiple such indicators, another would be to direct such essays to a human rater for review.

2.2.2 Gaming

> *Won't this grading system be easy to con? Can't the shrewd student just put in the proxies which will get a good grade?* (Page, 1966)

Certainly, students can and do employ gaming strategies to discover and exploit weaknesses of AWE systems. Such strategies can involve repeating the same paragraphs over and over, varying sentence structure, replacing words with more sophisticated variants, re-using words from the prompt, using general academic words, plagiarizing from other responses or from material found on the Internet, inserting unnecessary *shell language* – linguistic scaffolding for organizing claims and arguments, and automated generation of essays (Powers et al., 2001; Bejar et al., 2013, 2014; Higgins and Heilman, 2014; Sobel et al., 2014). Such strategies are generally handled by building in filters or flags for aberrant responses (Higgins et al., 2006; Zhang et al., 2016; Yoon et al., 2018; Cahill et al., 2018). However, developers of AWE systems can never anticipate all possible strategies and may have to react quickly as new ones are discovered in use, by developing new AWE methods to identify them. This cat-and-mouse game is particularly rampant in the context of standardized testing (§3.2). This is one of the reasons standardized tests are often not scored *solely* by an AWE system but also by a human rater.

2.2.3 Content

> *We are talking awfully casually about grading subject matter like history. Isn't this a wholly different sort of problem? Aren't we supposed to see that what the students are saying makes sense, above and beyond their using commas in the right places?* (Page, 1966)

Indeed, work has been done over the last decade on automated evaluation of written responses for their content and not their general writing quality (Sukkarieh and Bolge, 2008; Mohler et al., 2011; Ziai et al., 2012; Basu et al., 2013; Madnani et al., 2013; Ramachandran et al., 2015; Burrows et al., 2015; Sakaguchi et al., 2015; Madnani et al., 2016; Padó, 2016; Madnani et al., 2017a; Riordan et al., 2017; Kumar et al., 2017; Horbach et al., 2018; Riordan et al., 2019). Scoring for content focuses primarily on what students know, have learned, or can do in a *specific* subject area such as Computer Science, Biology, or Music, with the fluency of the response being secondary. For example, some spelling or grammar errors are acceptable as long as the desired specific information (e.g., scientific principles, trends in a graph, or details from a reading passage) is included in the response. Note that most current content scoring systems ascertain the "correctness" of a response based on its similarity to other responses that humans have deemed to be correct or, at least, high-scoring; they do not employ explicit fact-checking or reasoning for this purpose.

Concerns about specific content extends to other cases where the scoring system needs to pay attention to details of genre and task – not all essays are five-paragraph persuasive essays; the specific task might require assessing whether the student has appropriately used specific source materials (Beigman Klebanov et al., 2014; Rahimi et al., 2017; Zhang and Litman, 2018) or assessing narrative (Somasundaran et al., 2018) or reflective (Beigman Klebanov et al., 2016a; Luo and Litman, 2016), rather than persuasive, writing.

2.2.4 Feedback

Page emphasized the importance of feedback, and considered the following to be "the sort of feedback that can *almost* be programmed right now" (original italics):

> *John [. . .], please correct the following misspellings: believe, receive. Note the ie/ei problem. You overuse the words interesting, good, nice; then was repeated six times. Check trite expressions. All of your sentences are of the subject-verb variety and all are declarative. Reconstruct. Check subject-verb agreement in second paragraph. You had trouble with this in your last paper. Title lacking. Do the following related assignments for tomorrow . . .* (Page, 1966)

Today a substantial amount of writing feedback, particularly about spelling and grammar, is incorporated into widely used text editors such as Microsoft Word, Google Docs, and Overleaf. Dedicated writing assistance software such as ETS's Writing Mentor®[5] (Burstein et al., 2018), ASU's Writing Pal[6] (Roscoe and McNamara, 2013; Allen et al., 2014), ETS' *Criterion*®[7] (Burstein et al., 2004), Grammarly's Writing Assistant,[8] CambridgeEnglish's Write & Improve,[9] Ginger's Essay Checker,[10] TurnItIn's Revision Assistant,[11] Vantage Learning's MY Access!,[12] Pearson's My Writing Lab Writing Practice Module and WriteToLearn™[13,14] typically go beyond grammar and spelling.[15] Such tools provide feedback on discourse structure (Criterion), topic development and coherence (Writing Mentor), tone (Writing Assistant, Rao and Tetreault (2018)), thesis relevance (Writing Pal), sentence "spicing" through suggestions of synonyms and idioms (Ginger's Sentence Rephraser), and style & argumentation-related feedback (Revision Assistant).

Can we then put a green check-mark against Page's agenda for automated feedback, which "may magnify and disseminate the best human capacities to criticize, evaluate, and correct"? Alas, not yet; research on effectiveness of automated feedback on writing is inconclusive (Englert et al., 2007; Shermis et al., 2008; Grimes and Warschauer, 2010; Choi, 2010; Roscoe and McNamara, 2013; Wilson and Czik, 2016; Wilson, 2017; Bai and Hu, 2017; Ranalli et al., 2017). One potential reason for the different outcomes is difference in user populations – feedback that works for L1 writers might not work for L2 writers; differences in ages, skill levels, presence or absence of learning disabilities could all play a role. Adjustment of the evaluation methodology to the specific *purpose* of the writing assistance tool is another issue for consideration; we will return to this issue in §4.

3 Going off the Page

So far, Page's outline of the promises and challenges of AWE have provided a good framework for surveying the field. There are also a number of developments that were not mapped on Page's chart; we turn to reviewing those next.

3.1 Assessing writing in multiple languages

In order to advance the work on understanding and assessing writing quality, there is clearly a need for a multi-lingual perspective, since methods developed for one language or dialect may not work for another. This consideration does not appear in Page (1966), yet it is an active line of subsequent work. While most of the research we cited so far has been on English, various aspects of writing evaluation, e.g., annotation, detection of various types of errors, and building AWE systems, have been researched for a variety of languages: Song et al. (2016), Rao et al. (2017), Shiue et al. (2017) worked with data in Chinese,

[5] https://mentormywriting.org/
[6] http://www.adaptiveliteracy.com/writing-pal
[7] http://www.ets.org/criterion
[8] https://www.grammarly.com/
[9] https://writeandimprove.com/
[10] https://www.gingersoftware.com/essay-checker
[11] https://www.turnitin.com/products/revision-assistant
[12] http://www.vantagelearning.com/products/my-access-school-edition/
[13] https://www.pearsonmylabandmastering.com
[14] http://wtl.pearsonkt.com
[15] Writing Pal does not provide specific grammar and spelling feedback.

Lorenzen et al. (2019) in Danish, Berggren et al. (2019) in Norwegian, Amorim and Veloso (2017) in Portuguese, Stymne et al. (2017) in Swedish, Berkling (2018) and Weiss and Meurers (2019) in German, Mezher and Omar (2016) in Arabic, Kakkonen et al. (2005) in Finnish, Loraksa and Peachavanish (2007) in Thai, Lemaire and Dessus (2001) in French, and Ishioka and Kameda (2006) in Japanese. The list is by no means exhaustive; see Flor and Cahill (2020) for a recent review.

3.2 Standardized Testing

The use of automated evaluation technology envisioned by Page was as a service to reduce a teacher's burden; to eventually "lift from the shoulders of the English teacher, that brave and harried soul, his perpetual pressure of unassigned papers, or his unassuaged guilt." While such use has certainly been made (Burstein et al., 2004; Grimes and Warschauer, 2010), the most visible use case for AWE technology has arguably evolved to be in the context of standardized testing, be it for a test of English such as TOEFL® or PTE, a broader, more advanced psychometric examination such as the GRE® or GMAT, or for professional licensure such as AICPA or PRAXIS®.

This development of often high-stakes usage has led to somewhat different challenges from those that Page had anticipated. These challenges generally fall under the purview of the field of educational measurement (Bennett and Bejar, 1998; Clauser et al., 2002; Williamson et al., 2012): How to ensure that the automatic scores assigned to test takers are (1) *valid*, i.e., they actually measure the skill that the test developer designed the test to measure, (2) *defensible*, i.e., there is a reasonably clear explanation of why test takers received the particular scores they did, and (3) *fair* to all the test takers. We address each of these challenges separately below. Note that an additional challenge of high-stakes usage, not elaborated on here, is how to architect scoring systems for large-scale, low-latency use which requires them to be reliable, scalable, flexible, and attentive to the choice of software and application frameworks (Madnani et al., 2018).

3.2.1 Construct Validity

Page declares that he is not after "generating measures of what the true characteristics of the essays are, as ordinarily discussed by human raters" but rather is content "to settle for the correlates of these true characteristics." Page seems to do away rather quickly with trying to measure the actual thing – the set of all and only "true characteristics of essays", or *trins*. Why is that? He explains:

> *Notwithstanding the wonders of the computer, we have to develop a strategy in order to tell the computer what to do. The difficult part is the development of this strategy. It is difficult because we do not really understand what the psychological components are in the judgment of essays. It is easy enough to get persons to expound authoritatively on such judgment, but the fuzziness and inutility of their thinking becomes at once evident when the effort is made to translate it into a computer program.* (Page, 1966)

Page's argument is that we do not know precisely enough what the human raters are doing to try and implement that. Some work on rater cognition has already been done in the early 1950s and 1960s, e.g., in the context of the College Entrance Examination Board's development of the General Composition Test. Diederich et al. (1961) had 53 distinguished individuals from various academic disciplines and beyond (English, Social Science, Natural Science, Law, Writers and Editors, Business Executives) sort student essays "in order of merit", with no definition thereof, instructing readers as follows:

> *Use your own judgment as to what constitutes "writing ability." Do not assume that we want you to do this or that. We want you to use whatever hunches, intuitions, or preferences you normally use in deciding that one piece of writing is better than another. You need not even act as a representative of your field, since individuals in any field have varying tastes and standards.*

Readers were also asked to a write brief comments on anything that they liked or disliked about the essay, on as many essays as possible. For the study, a sample of U.S. college freshmen were asked to write essays in response to four topics as part of homework. A total of 300 essays addressing two topics were chosen for the analyses, sampled so as to make sure that the full range of abilities is represented (approximated via SAT Verbal scores). The researchers performed a factor analysis on the matrix of pairwise correlations among the readers, and identified groups of readers (factors) that represent five "schools of thought" about writing quality. Analyzing the comments made by readers who belong to the different "schools of thought", they identified five categories that were each prioritized by one of the groups of readers:

1. Ideas (including relevance, clarity, quantity, development, persuasiveness)
2. Form (including spelling, organization, analysis, coherence)
3. Flavor (including style, originality, quality of ideas, interest, sincerity)
4. Mechanics (including punctuation, grammar, sentence structure, phrasing)
5. Wording (including felicity of expression, comments on specific word choices, cliches)

It is based on such findings above that general scoring criteria have emerged (Deane, 2013) and morphed into *scoring rubrics*. These are explicit criteria set by and for human raters for evaluating essays. For example, to score highly on the GRE® Issue essay-writing task,[16] one typically:

- articulates a clear and insightful position on the issue in accordance with the assigned task
- develops the position fully with compelling reasons and/or persuasive examples
- sustains a well-focused, well-organized analysis, connecting ideas logically
- conveys ideas fluently and precisely, using effective vocabulary and sentence variety
- demonstrates superior facility with the conventions of standard written English (i.e., grammar, usage and mechanics), but may have minor errors

In the current practice of automated scoring of standardized tests, developers of a scoring engine often need to provide a *construct validity argument* in order to show that what the system is measuring is actually aligned with the "writing construct" – the actual set of writing skills that the test is supposed to measure.

[16] https://www.ets.org/gre/revised_general/prepare/analytical_writing/issue/scoring_guide

Some of the items in a human-oriented scoring rubrics are amenable to reasonably direct implementation, often with the help of human-annotated gold standard data such as misspellings (Flor, 2012; Flor and Futagi, 2013) and specific grammar errors (Rozovskaya and Roth, 2010; Leacock et al., 2014). It might be the case that the system would miss some grammar errors and declare an error where there is none, but a grammar assessment system can be built for identifying specific, observable instances of errors that a human reader focused on Mechanics would likely pick upon.

For other items in a rubric, one might need to drill down, articulate a reliable guideline for humans to assess that particular aspect of the essay, annotate a substantial enough number of essays using the guidelines to make machine learning possible, and then find automatically measurable properties of essays that would provide information relevant to that particular aspect of essay quality. This would be a mix between what Page called a *prox* and a *trin*, in that a particular, intrinsically interesting, aspect of an essay can be identified reliably by humans, and an automated system can learn how to approximate that particular construct. Such approaches have been developed for organization (well-organized) (Burstein et al., 2003), coherence (well-focused, conveys ideas fluently) (Burstein et al., 2010; Somasundaran et al., 2014), grammaticality (facility with conventions) (Heilman et al., 2014), thesis clarity (clarity) (Persing and Ng, 2013) as well as aspects of scoring rubrics that are more task-specific, e.g., argumentation (clear position, with compelling reasons) (Stab and Gurevych, 2014; Ghosh et al., 2016; Beigman Klebanov et al., 2017; Stab and Gurevych, 2017; Carlile et al., 2018), use of evidence in the context of source-based writing (Rahimi et al., 2017).

Finally, for some rubric items, it is not clear exactly how to reliably translate the relevant aspect of the writing construct into annotations guidelines, and so *proxes* might be employed. For example, consider Page's argument for capturing "diction" (appropriate word choice) through word frequency – a writer who can use many different words, including rarer and often semantically nuanced ones, is likelier to make precise word choices than a writer who uses a more limited vocabulary. Attempts to capture topicality (Beigman Klebanov et al., 2016b) or development

(Beigman Klebanov and Flor, 2013b; Somasundaran et al., 2016) through properties of vocabulary distribution without human annotation of topicality and development exemplify such approaches.

3.2.2 Model Interpretability

Recent research has shown that more sophisticated machine learning models might perform better than simple regression-based models when it comes to predictive accuracy (Chen and He, 2013; Cummins et al., 2016; Taghipour and Ng, 2016; Alikaniotis et al., 2016; Dong et al., 2017; Dasgupta et al., 2018; Jin et al., 2018). However, unlike linear regression where stakeholders can understand how much each feature used in the model contributed to the predicted score, many of the more complex models are essentially "black boxes" and do not really lend themselves to *post-hoc interpretability* (Lipton, 2016). Although interpretability is an active area of research in the machine learning literature (Ribeiro et al., 2016; Koh and Liang, 2017; Doshi-Velez and Kim, 2017), it currently lags behind the research on machine learning methods. For this reason, some automated scoring systems used for high-stakes standardized testing – like ETS's e-Rater (Attali and Burstein, 2006) – still use some variant of least squares linear regression as the machine learning model to predict test taker scores.

3.3 Increased Attention to Fairness

It would probably not be an overstatement to say that fairness in AI is quickly becoming its own sub-field, with a new annual ACM conference on Fairness, Accountability, and Transparency having been inaugurated in 2018[17] and relevant research appearing at many impactful publication venues, such as Science (Caliskan et al., 2017), NIPS (Pleiss et al., 2017; Kim et al., 2018), ICML (Kearns et al., 2018), ACL (Hovy and Spruit, 2016; Sun et al., 2019; Sap et al., 2019), KDD (Speicher et al., 2018), AAAI (Zhang and Bareinboim, 2018), and others (Dwork et al., 2012; Hajian and Domingo-Ferrer, 2013). There is also recent work that examines fairness and ethical considerations when using AI in an education (Mayfield et al., 2019; Gardner et al., 2019).

In the context of assessment, fairness considerations dictate that the test reflects the same construct(s) for the entire test taking population, that scores from the test have the same meaning for all the test taking population, and that a fair test does not offer undue advantages (or disadvantages) to some individuals because of their characteristics – such as those associated with race, ethnicity, gender, age, socioeconomic status, or linguistic or cultural background – or the test characteristics itself, e.g., the different prompts shown to different test-takers at test time.

The educational measurement community has long been studying fairness in automated scoring (Williamson et al., 2012; Ramineni and Williamson, 2013; AERA, 2014) and recent progress made by the NLP community towards enhancing the usual accuracy-based evaluations with some of these psychometric analyses – from computing indicators of potential biases in automatic scores across various demographic sub-groups to computing new metrics that incorporate measurement theory to produce more reliable indicators of system performance – is quite promising (Madnani et al., 2017b; Loukina et al., 2019).

3.4 Pervasiveness of Technology

Page's *gedankenexperiment* on the potential of automated essay evaluation in a classroom context no doubt appeared audacious in 1966 but nothing back then could have prepared his readers to the pervasiveness of technology we are experiencing today. Today you can very literally carry your AWE system in your pocket; you can even carry several. You can use them (almost) at any time and at any place – not only in classrooms, but at home, at work, and even while texting with a friend.

This is perhaps the biggest issue that Page's vision did not address: the possibility of universal availability and the concomitant co-optation of a tool beyond its original intended purpose. Much like the calculator – invented by Blaise Pascal to help his father with the tedious arithmetic of tax collection – ended up "freeing" people from the burden of figuring out their intended tip at a restaurant through mental arithmetic, a future writing aid meant to help a student improve his argument writing assignment for a class could end up being used by a lawyer for composing his closing argument. Since such usages are on the horizon, we should consider the implications now.

[17] https://facctconference.org/

4 Discussion

Once an invention is out in the open, it is difficult to predict what specific uses people would put it to. How do we go about evaluating the tool if we don't know what the user's goal is? While it isn't possible to anticipate all specific uses, it is possible, we believe, to consider the *types* of uses that suggest different evaluation strategies. From the current vantage point, we see three types of uses.

4.1 Support Consequential Decision Making

The first use is where a consequential decision about the writer or a related entity (such as a class or a school) is being made based on the written product. This use is exemplified by the application of automated scoring in a standardized testing context to decide on admissions to an institution of higher education or the granting of a professional licenses; other cases such as course placement decisions, coursework grading, or even extension of a job offer (where the submission of a writing sample is a part of the job application process) would belong to this type of use. In all such cases, the automated system needs to provide valid and fair scores (or other types of feedback), since the livelihood or professional trajectory of people might depend on the outcome. We have dealt with the particulars of this case in detail in §3.2.

4.2 Create a Better Written Product

The second type of use is one where the focus is on the final product, namely, the actual piece of writing produced following the writer's use of AWE technology. In this context, it does not much matter exactly what part of the final product is due to the human and which part is due to the machine – perhaps the machine only corrected misspellings, or suggested improvements for the human to vet, or maybe the human only contributed the very first ideation, and the machine has done the rest. Perhaps all the human writer contributed was the thesis ('I think school should start at 8 rather than 7') and then clicked 'submit' to get back an essay making a cogent and convincing case in support of the thesis. Mining large textual databases for arguments and evaluating them are feasible today as recently demonstrated by IBM's Debater technology[18] (Rinott et al., 2015; Levy et al., 2017; Gretz et al., 2019); introduce some figuration to make it more appealing (Veale et al., 2017; Veale, 2018) and storify it (Riegl and Veale, 2018; Radford et al., 2019), et voilà!

This type of use is essentially a machine's augmentation of human ability, and is hinted at, for example, in a customer testimonial for Grammarly: "Grammarly allows me to get those communications out and feel confident that I'm putting my best foot forward. Grammarly is like a little superpower, especially when I need to be at 110%." The human presumably remains at the same level of ability, but the product of the machine-human collaboration is superior to what the human alone could have produced.

In this context, the primary evaluation criterion for AWE is the fitness of the resulting communication to its purpose, or, at least, some evidence of improvement of the product over the human's first draft. Indeed, measurements of improvement across drafts and evidence of students' making corrections following feedback are often used for evaluation (Attali, 2004; Lipnevich and Smith, 2008; Foltz et al., 2014; Chapelle et al., 2015).

Within the product-centered evaluation paradigm, there could be various specific objectives other than the improvement of the holistic quality of the piece of writing; it could be an increase in the speed of production, or the maximization of click-through rate in an advertisement text, for example.

4.3 Help the User Learn to Write Better

The third type of use for AWE software is to help the writer improve his or her writing skill. Scores or other types of feedback are designed, in this context, to provide tutoring or guidance, not for fixing specific problems in the current piece of writing but to help the user learn more general skills that would make the first draft of their *next* essay better than the first draft of their current essay.

Evaluation of a tool though a demonstration of skill-improvement – the efficacy of the tool – is a complicated endeavor. To demonstrate that the observed improvement in skill is specifically due to the use of the writing tool, and not due to something else happening in students' life and education at the same time requires a research design that can take other potential sources of variation in outcomes into account, such as the one used in randomized controlled studies often used to as-

[18] https://www.research.ibm.com/artificial-intelligence/project-debater/

sess interventions, including in education (Connolly et al., 2018); some such studies have been performed with respect to AWE tools (Rock, 2007; Wilson and Roscoe, 2020). A tool that allows for monitoring of improvement in skill (even if the improvement is due to other factors such as school instruction or participation in some activity or community) could also be useful in the broader context of skill-oriented use, as the learner and the teacher would be able to tell that improvement is happening, even if we do not know exactly why. Improvement in important aspects of learning such as motivation and self-efficacy could also provide value to the learner (Grimes and Warschauer, 2010; Wilson and Roscoe, 2020).

4.4 Relationships between Types of Use

One could argue that an ideal automated writing assistant would support all the different goals at once – help one produce better writing, help one learn, and do both in a psychometrically responsible fashion – benefits are not restricted to certain types of users more than others – so that decision-making based on the outcome of the usage of the tool can also be supported.

Indeed, the uses are not necessarily mutually exclusive. For example, the human augmentation and consequential decision use cases could apply at the same time. It is possible that, at some future point in time, spelling will be deemed to lie outside of the construct targeted by the consequential assessment of writing and spell-correction software will be made available to test-takers. However, this would require a careful examination of the impact of correction on the distributions and interpretations of the scores. In particular, Choi and Cho (2018) found that manually-vetted correction of spelling errors yielded a significant increase in scores assigned to the essays by trained raters, and that, even after controlling for the error quantity and quality predictors, the magnitude of the average gain in the score was smaller for responses with higher original scores. Add to the mix the finding that automated spelling correction system is more accurate on essays that are of better quality to begin with (Flor, 2012), and it's likely that the automated assessment of an automatically spell-corrected version of an essay might show an unexpected relationship with original scores that would need to be closely examined for bias or for an increase in construct-irrelevant variance.

It is also possible that the effect of using a tool optimized for one use case could be the opposite of what another use case requires. If 'use it or lose it' has any truth to it, a potential consequence of extensive, consistent, and pervasive human augmentation for producing superior written products is an adverse impact on the skill of the human in the human-machine team. If the near universal adoption of calculators is any guide, once a skill (long division) can be reliably outsourced to a machine, humans stop valuing it in daily practice and, therefore, might set out to lose it in the long run.[19] Spelling is a likely candidate writing skill where reliable access to high quality correction software could make humans stop worrying about it rather than invest effort in improving it.

Many of the tools mentioned in §2.2.4 seem to position themselves somewhere between the skill-improvement and the product-improvement use cases, perhaps assuming that quantity will eventually turn into quality, namely, extensive work on improving the written product might lead to internalization and generalization of the skill to new contexts. This might or might not be true. Feedback that helps the user fix an error quickly by pointing it out and by suggesting a correction might be good in a product-oriented context, but not in a skill-oriented context; letting the user pinpoint and fix the error himself or herself might be a better skill-development strategy (Hyland and Hyland, 2006). According to Graham and Perin (2007) meta-analysis of writing interventions for adolescents, explicit grammar instruction tended to be ineffective; this finding is cited by the developers for Writing Pal to support their decision to forgo giving explicit feedback on grammar (McNamara et al., 2013), in contrast to most other AWE systems that do provide such feedback.

5 Summary & Conclusion

In his visionary paper from 1966, Ellis Page provided a proof-of-concept demonstration of the possibility of automated grading of essays, as well

[19] 1989 Curriculum and Evaluation Standards for School Mathematics from the National Council of Teachers of Mathematics recommend in the Summary of Changes to Content and Emphasis in K-4 Mathematics (p.21) decreasing the attention devoted to long division specifically and to "complex paper-and-pencil computations" in general; the recommendation for grades 5-8 is likewise to decrease emphasis on "tedious paper-and-pencil computations" (p.71). https://archive.org/details/curriculumevalua00nati. The document has sparked substantial controversy, including with regards to long division (Klein and Milgram, 2000).

as outlined some potential challenges to its adoption. Subsequent research and practice have delivered on Page's minimum desiderata for an AWE system; current research is working to address the outstanding challenges dealing with a variety of languages, content domains, and writing tasks.

The field of AWE has thus progressed according to the trajectory charted by Page to a large extent, though not completely. In particular, while Page imagined the main use case of AWE to be in the service of a harried English teacher and his feedback-thirsty students, in reality, the most visible use case has arguably evolved to be automated scoring of essays for standardized testing, which, in turn, has led to new challenges, such as ensuring the validity and fairness of scores.

The other development that Page could not anticipate is the sheer pervasiveness of technology in people's daily lives; AWE software can be made available not only in classrooms to be used under the watchful eye of the English teacher, but (almost) anywhere and at any time, including on mobile devices. While it is difficult to predict specific uses people would find for such software, we outlined a number of *types* of use, depending on the goal: (a) consequential decision making about the user; (b) delivery of the best possible written product in partnership with the user; and (c) assisting the user in improving her writing skills. We believe that we, as researchers, can help users find *value* in our technology by considering the goals, engaging partners from other relevant disciplines, and designing the tools as well as their e*valu*ations to focus on specific types of use.

Acknowledgements

We would like to thank our colleagues Anastassia Loukina, Jill Burstein, Aoife Cahill, and Isaac Bejar, as well as ACL reviewers and area chair, for their thoughtful comments on earlier drafts of this paper.

References

AERA. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725.

Laura Allen, Scott Crossley, Erica Snow, and Danielle McNamara. 2014. L2 writing practice: Game enjoyment as a key to engagement. *Language Learning & Technology*, 18(2):124–150.

Evelin Amorim and Adriano Veloso. 2017. A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102.

Yigal Attali. 2004. Exploring the feedback and revision features of criterion. *Journal of Second Language Writing*, 14:191–205.

Yigal Attali and Jill Burstein. 2006. Automated Essay Scoring with e-raterő V. 2. *The Journal of Technology, Learning and Assessment*, 4(3):1–30.

Lifang Bai and Guangwei Hu. 2017. In the face of fallible awe feedback: how do students respond? *Educational Psychology*, 37(1):67–81.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a Clustering Aapproach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Beata Beigman Klebanov, Jill Burstein, Judith Harackiewicz, Stacy Priniski, and Matthew Mulholland. 2016a. Enhancing STEM motivation through personal and communal values: NLP for assessment of utility value in student writing. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–205.

Beata Beigman Klebanov and Michael Flor. 2013a. Argumentation-relevant metaphors in test-taker essays. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 11–20.

Beata Beigman Klebanov and Michael Flor. 2013b. Word association profiles and their use for automated scoring of essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1148–1158.

Beata Beigman Klebanov, Michael Flor, and Binod Gyawali. 2016b. Topicality-Based Indices for Essay Scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 63–72.

Beata Beigman Klebanov, Binod Gyawali, and Yi Song. 2017. Detecting Good Arguments in a Non-Topic-Specific Way: An Oxymoron? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 244–249.

Beata Beigman Klebanov, Chee Wee (Ben) Leong, and Michael Flor. 2018. A corpus of non-native written English annotated for metaphor. In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 86–91.

Beata Beigman Klebanov, Nitin Madnani, Jill Burstein, and Swapna Somasundaran. 2014. Content importance models for scoring writing from sources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 247–252.

Isaac Bejar, Michael Flor, Yoko Futagi, and Chaintanya Ramineni. 2014. On the vulnerability of automated scoring to construct-irrelevant response strategies (cirs): An illustration. *Assessing Writing*, 22:48–59.

Isaac Bejar, Waverely VanWinkle, Nitin Madnani, William Lewis, and Michael Steier. 2013. Length of textual response as a construct-irrelevant response strategy: The case of shell language. *ETS Research Report Series*, 2013(1):1–39.

Randy Elliot Bennett and Isaac I Bejar. 1998. Validity and automad scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 17(4):9–17.

Stig Johan Berggren, Taraka Rama, and Lilja Øvrelid. 2019. Regression or classification? automated essay scoring for Norwegian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 92–102.

Kay Berkling. 2018. A 2nd longitudinal corpus for childrens writing with enhanced output for specific spelling patterns. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, 25(1):60–117.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2004. Automated essay evaluation: The criterion online writing service. *AI Magazine*, 25(3):27–36.

Jill Burstein, Norbert Elliot, Beata Beigman Klebanov, Nitin Madnani, Diane Napolitano, Max Schwartz, Patrick Houghton, and Hilary Molloy. 2018. Writing mentor: Writing progress using self-regulated writing support. *Journal of Writing Analytics*, 2:314–328.

Jill Burstein, Daniel Marcu, and Kevin Knight. 2003. Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems*, 18(1):32–39.

Jill Burstein, Joel Tetreault, and Slava Andreyev. 2010. Using Entity-Based Features to Model Coherence in Student Essays. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 681–684.

Aoife Cahill, Martin Chodorow, and Michael Flor. 2018. Developing an e-rater Advisory to Detect Babel-generated Essays. *Journal of Writing Analytics*, 2:203–224.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631.

Carol Chapelle, Elena Cotos, and Jooyoung Lee. 2015. Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, 32(3):385–405.

Hongbo Chen and Ben He. 2013. Automated Essay Scoring by Maximizing Human-Machine Agreement. In *Proceedings of EMNLP*, pages 1741–1752.

Ikkyu Choi and Yeonsuk Cho. 2018. The impact of spelling errors on trained raters' scoring decisions. *Language Education & Assessment*, 1(2):45–58.

Jaeho Choi. 2010. *The Impact of Automated Essay Scoring (AES) for Improving English Language Learner's Essay Writing*. University of Virginia Charlottesville, VA.

Brian E. Clauser, Michael T. Kane, and David B. Swanson. 2002. Validity issues for performance-based tests scored with computer-automated scoring systems. *Applied Measurement in Education*, 15(4):413–432.

Paul Connolly, Ciara Keenan, and Karolina Urbanska. 2018. The trials of evidence-based practice in education: A systematic review of randomised controlled trials in education research 1980-2016. *Educational Research*.

Ronan Cummins, Meng Zhang, and Ted Briscoe. 2016. Constrained Multi-Task Learning for Automated Essay Scoring. In *Proceedings of ACL*, pages 789–799.

Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. 2018. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102.

Paul Deane. 2013. On the Relation between Automated Essay Scoring and Modern Views of the Writing Construct. *Assessing Writing*, 18(1):7–24.

Paul Diederich, John French, and Sydell Carlton. 1961. Factors in judgments of writing ability. *ETS Research Bulletin*, RB-61-15.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162.

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *CoRR*, abs/1702.08608.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS 12, page 214226.

Carol Englert, Yong Zhao, Kailonnie Dunsmore, Natalia Yevgenyevna Collings, and Kimberly Wolbers. 2007. Scaffolding the writing of students with disabilities through procedural facilitation: Using an internet-based technology to improve performance. *Learning Disability Quarterly*, 30(1):9–29.

Michael Flor. 2012. Four types of context for automatic spelling correction. *Traitement Automatique des Langues (TAL)*, 53(3):61–99.

Michael Flor and Aoife Cahill. 2020. Automated scoring of open-ended written responses – possibilities and challenges. In *Innovative Computer-based International Large-Scale Assessments*. Springer Science Publishers.

Michael Flor and Yoko Futagi. 2013. Producing an annotated corpus with automatic spelling correction. In S. Granger, G. Gilquin, and F. Meunier, editors, *Twenty Years of Learner Corpus Research: Looking back, moving ahead*, pages 139–154. Louvain-la-Neuve: Presses universitaires de Louvain.

Peter Foltz, Mark Rosenstein, Nicholas Dronen, and Scott Dooley. 2014. Automated feedback in a large-scale implementation of a formative writing system: Implications for improving student writing. In *Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA*.

Josh Gardner, Christopher Brooks, and Ryan Baker. 2019. Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, LAK19, page 225234.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554.

Steve Graham and Dolores Perin. 2007. A meta-analysis of writing instruction for adolescent students. *Journal of educational psychology*, 99(3):445.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis.

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

S. Hajian and J. Domingo-Ferrer. 2013. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445–1459.

Michael Heilman, Aoife Cahill, Nitin Madnani, Melissa Lopez, Matthew Mulholland, and Joel Tetreault. 2014. Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 174–180.

Derrick Higgins, Jill Burstein, and Yigal Attali. 2006. Identifying Off-topic Student Essays without Topic-specific Training Data. *Natural Language Engineering*, 12(2):145–159.

Derrick Higgins and Michael Heilman. 2014. Managing what we can measure: Quantifying the susceptibility of automated scoring systems to gaming behavior. *Educational Measurement: Issues and Practice*, 33(3):36–46.

Andrea Horbach, Sebastian Stennmanns, and Torsten Zesch. 2018. Cross-Lingual Content Scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 410–419.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

Ken Hyland and Fiona Hyland. 2006. Feedback on second language students' writing. *Language teaching*, 39(2):83–101.

Tsunenori Ishioka and Masayuki Kameda. 2006. Automated japanese essay scoring system based on articles written by experts. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 233–240. Association for Computational Linguistics.

Cancan Jin, Ben He, Kai Hui, and Le Sun. 2018. TDNN: A Two-Stage Deep Neural Network for Prompt-Independent Automated Essay Scoring. In *Proceedings of ACL*, pages 1088–1097.

Tuomo Kakkonen, Niko Myller, Jari Timonen, and Erkki Sutinen. 2005. Automatic essay grading with probabilistic latent semantic analysis. In *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, EdAppsNLP 05, pages 29–36.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: a survey of the state of the art. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6300–6308. AAAI Press.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572.

Michael Kim, Omer Reingold, and Guy Rothblum. 2018. Fairness through computationally-bounded awareness. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4842–4852. Curran Associates, Inc.

David Klein and James Milgram. 2000. The role of long division in the K–12 curriculum. `https://www.csun.edu/~vcmth00m/longdivision.pdf`.

Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894.

Sachin Kumar, Soumen Chakrabarti, and Shourya Roy. 2017. Earth mover's distance pooling over siamese lstms for automatic short answer grading. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, pages 2046–2052.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2014. Automated grammatical error detection for language learners. *Synthesis lectures on human language technologies*, 7(1):1–170.

Benoit Lemaire and Philippe Dessus. 2001. A system to assess the semantic content of student essays. *Journal of Educational Computing Research*, 24(3):305–320.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining*, pages 79–84.

Anastasiya Lipnevich and Jeffrey Smith. 2008. Response to assessment feedback: The effect of grades, praise, and source of information. *ETS Research Report No. RR-08-30*.

Zachary Lipton. 2016. The Mythos of Model Interpretability. In *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*.

Jeannette Littlemore, Tina Krennmayr, James Turner, and Sarah Turner. 2014. An investigation into metaphor use at different levels of second language writing. *Applied Linguistics*, 35(2):117–144.

Chanunya Loraksa and Ratchata Peachavanish. 2007. Automatic Thai-language essay scoring using neural network and latent semantic analysis. In *First Asia International Conference on Modelling Simulation (AMS'07)*, pages 400–402.

Stephan Lorenzen, Niklas Hjuler, and Stephen Alstrup. 2019. Investigating writing style development in high school. *CoRR*, abs/1906.03072.

Annie Louis and Ani Nenkova. 2013. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the Association for Computational Linguistics*, 1:341–352.

Anastassia Loukina, Nitin Madnani, and Klaus Zechner. 2019. The Many Dimensions of Algorithmic Fairness in Educational Applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10.

Wencan Luo and Diane Litman. 2016. Determining the quality of a student reflective response. In *Proceedings of the Florida Artificial Intelligence Research Society Conference*.

Nitin Madnani, Jill Burstein, John Sabatini, and Tenaha O'Reilly. 2013. Automated Scoring of a Summary-Writing Task Designed to Measure Reading Comprehension. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–168.

Nitin Madnani, Aoife Cahill, Daniel Blanchard, Slava Andreyev, Diane Napolitano, Binod Gyawali, Michael Heilman, Chong Min Lee, Chee Wee Leong, Matthew Mulholland, and Brian Riordan. 2018. A robust microservice architecture for scaling automated scoring applications. *ETS Research Report Series*, 2018(1).

Nitin Madnani, Aoife Cahill, and Brian Riordan. 2016. Automatically scoring tests of proficiency in music instruction. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 217–222.

Nitin Madnani, Anastassia Loukina, and Aoife Cahill. 2017a. A Large Scale Quantitative Exploration of Modeling Strategies for Content Scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 457–467.

Nitin Madnani, Anastassia Loukina, Alina von Davier, Jill Burstein, and Aoife Cahill. 2017b. Building Better Open-Source Tools to Support Fairness in Automated Scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52.

Elijah Mayfield, Michael Madaio, Shrimai Prabhumoye, David Gerritsen, Brittany McLaughlin, Ezekiel Dixon-Román, and Alan W Black. 2019. Equity beyond bias in language technologies for education. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 444–460.

Danielle McNamara, Scott Crossley, and Rod Roscoe. 2013. Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, 45(2):499–515.

Rabih Mezher and Nazlia Omar. 2016. A hybrid method of syntactic feature and latent semantic analysis for automatic arabic essay scoring. *Journal of Applied Sciences*, 16(5):209.

Michael Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In *Proceedings of ACL: HLT*, pages 752–762.

Ulrike Padó. 2016. Get Semantic With Me! The Usefulness of Different Feature Types for Short-Answer Grading. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2186–2195.

Ellis B Page. 1966. The Imminence of Grading Essays by Computer. *The Phi Delta Kappan*, 47(5):238–243.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689.

Donald Powers, Jill Burstein, Martin Chodorow, Mary Fowles, and Karen Kukich. 2001. Stumping E-rater: Challenging the Validity of Automated Essay Scoring. *ETS Research Report Series*, 2001(1):1–44.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Zahra Rahimi, Diane Litman, Richard Correnti, Elaine Wang, and Lindsay Clare Matsumura. 2017. Assessing students use of evidence and organization in response-to-text writing: Using natural language processing for rubric-based automated scoring. *International Journal of Artificial Intelligence in Education*, 27(4):694–728.

Lakshmi Ramachandran, Jian Cheng, and Peter Foltz. 2015. Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications*, pages 97–106, Denver, Colorado.

Chaitanya Ramineni and David Williamson. 2013. Automated essay scoring: Psychometric guidelines and practices. *Assessing Writing*, 18(1):25 – 39. Automated Assessment of Writing.

Jim Ranalli, Stephanie Link, and Evgeny Chukharev-Hudilainen. 2017. Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology*, 37(1):8–25.

Gaoqi Rao, Baolin Zhang, Endong Xun, and Lung-Hao Lee. 2017. Ijcnlp-2017 task 1: Chinese grammatical error diagnosis. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 1–8.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140, New Orleans, Louisiana. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Stefan Riegl and Tony Veale. 2018. Live, die, evaluate, repeat: Do-over simulation in the generation of coherent episodic stories. In *Proceedings of the Ninth International Conference on Computational Creativity*, pages 80–87, Salamanca, Spain.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence-an automatic method for context dependent evidence detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 440–450.

Brian Riordan, Michael Flor, and Robert Pugh. 2019. How to Account for Mispellings: Quantifying the Benefit of Character Representations in Neural Content Scoring Models. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 116–126, Florence, Italy.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. 2017. Investigating Neural Architectures for Short Answer Scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168.

JoAnn Leah Rock. 2007. The impact of short-term use of criterionsm on writing skills in ninth grade. *ETS Research Report Series*, 2007(1):i–24.

Rod D Roscoe and Danielle S McNamara. 2013. Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, 105(4):1010.

Alla Rozovskaya and Dan Roth. 2010. Annotating ESL errors: Challenges and rewards. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36.

Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective Feature Integration for Automated Short Answer Scoring. In *Proceedings of NAACL: HLT*, pages 1049–1054.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Mark Shermis, Cynthia Garvan, and Yanbo Diao. 2008. The impact of automated essay scoring on writing outcomes. *Online Submission, ERIC*.

Yow-Ting Shiue, Hen-Hsen Huang, and Hsin-Hsi Chen. 2017. Detection of Chinese word usage errors for non-native Chinese learners with bidirectional LSTM. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 404–410, Vancouver, Canada. Association for Computational Linguistics.

Louis Sobel, Milo Beckman, Damien Jiang, and Les Perelman. 2014. BABEL Generator. `http://babel-generator.herokuapp.com`.

Swapna Somasundaran, Jill Burstein, and Martin Chodorow. 2014. Lexical Chaining for Measuring Discourse Coherence Quality in Test-taker Essays. In *Proceedings of COLING*, pages 950–961.

Swapna Somasundaran, Michael Flor, Martin Chodorow, Hillary Molloy, Binod Gyawali, and Laura McCulla. 2018. Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6:91–106.

Swapna Somasundaran, Brian Riordan, Binod Gyawali, and Su-Youn Yoon. 2016. Evaluating argumentative and narrative essays using graphs. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1568–1578.

Wei Song, Tong Liu, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Learning to identify sentence parallelism in student essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 794–803.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD 18, pages 2239–2248.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 980–990.

Sara Stymne, Eva Pettersson, Beáta Megyesi, and Anne Palmér. 2017. Annotating errors in student texts: First experiences and experiments. In *Joint 6th NLP4CALL and 2nd NLP4LA Nodalida workshop*, pages 47–60.

Jana Sukkarieh and Eleanor Bolge. 2008. Leveraging c-raters automated scoring capability for providing instructional feedback for short constructed responses. In *Intelligent Tutoring Systems*, pages 779–783. Springer.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891.

Tony Veale. 2018. A massive sarcastic robot: What a great idea! two approaches to the computational generation of irony. In *Proceedings of the Ninth International Conference on Computational Creativity*, Salamanca, Spain.

Tony Veale, Hanyang Chen, and Guofu Li. 2017. I read the news today, oh boy - making metaphors topical, timely and humorously personal. In *Distributed, Ambient and Pervasive Interactions*, pages 696–709.

Zarah Weiss and Detmar Meurers. 2019. Analyzing linguistic complexity and accuracy in academic language development of german across elementary and secondary school. In *Proceedings of the*

Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 380–393.

David Williamson, Xiaoming Xi, and Jay Breyer. 2012. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Joshua Wilson. 2017. Associated effects of automated essay evaluation software on growth in writing quality for students with and without disabilities. *Reading and Writing*, 30(4):691–718.

Joshua Wilson and Amanda Czik. 2016. Automated essay evaluation software in english language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100(94-109).

Joshua Wilson and Rod Roscoe. 2020. Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1):87–125.

Su-Youn Yoon, Aoife Cahill, Anastassia Loukina, Klaus Zechner, Brian Riordan, and Nitin Madnani. 2018. Atypical Inputs in Educational Applications. In *Proceedings of NAACL*, pages 60–67.

Haoran Zhang and Diane Litman. 2018. Co-attention based neural network for source-dependent essay scoring. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 399–409.

Junzhe Zhang and Elias Bareinboim. 2018. Fairness in decision-makingthe causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Mo Zhang, Jing Chen, and Chunyi Ruan. 2016. Evaluating the Advisory Flags and Machine Scoring Difficulty in the e-rater® Automated Scoring Engine. *ETS Research Report Series*, 2016(2):1–14.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short Answer Assessment: Establishing Links Between Research Strands. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 190–200.