ETS TOEFL

Quality Beyond Measu

Check for updates

TOEFL® Research Report
TOEFL–RR-86
ETS Research Report No. RR–19-12

Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany

André A. Rupp

Jodi M. Casabianca

Maleika Krüger

Stefan Keller

Olaf Köller

December 2019

The *TOEFL*® test is the world's most widely respected English language assessment, used for admissions purposes in more than 130 countries including Australia, Canada, New Zealand, the United Kingdom, and the United States. Since its initial launch in 1964, the TOEFL test has undergone several major revisions motivated by advances in theories of language ability and changes in English teaching practices. The most recent revision, the *TOEFL iBT*® test, contains a number of innovative design features, including integrated tasks that engage multiple skills to simulate language use in academic settings and test materials that reflect the reading, listening, speaking, and writing demands of real-world academic environments. In addition to the TOEFL iBT, the TOEFL Family of Assessments has expanded to provide high-quality English proficiency assessments for a variety of academic uses and contexts. The TOEFL Young Students Series (YSS) features the *TOEFL*® *Primary*™ and *TOEFL Junior*® tests, designed to help teachers and learners of English in school settings. The *TOEFL ITP*® Assessment Series offers colleges, universities, and others an affordable test for placement and progress monitoring within English programs.

Since the 1970s, the TOEFL tests have had a rigorous, productive, and far-ranging research program. ETS has made the establishment of a strong research base a consistent feature of the development and evolution of the TOEFL tests, because only through a rigorous program of research can a testing company demonstrate its forward-looking vision and substantiate claims about what test takers know or can do based on their test scores. In addition to the 20-30 TOEFL-related research projects conducted by ETS Research & Development staff each year, the TOEFL Committee of Examiners (COE), composed of distinguished language-learning and testing experts from the academic community, funds an annual program of research supporting the TOEFL family of assessments, including projects carried out by external researchers from all over the world.

To date, hundreds of studies on the TOEFL tests have been published in refereed academic journals and books. In addition, more than 300 peer-reviewed reports about TOEFL research have been published by ETS. These publications have appeared in several different series historically: TOEFL Monographs, TOEFL Technical Reports, TOEFL iBT Research Reports, and TOEFL Junior Research Reports. It is the purpose of the current TOEFL Research Report Series to serve as the primary venue for all ETS publications on research conducted in relation to all members of the TOEFL Family of Assessments.

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org **Web site:** www.ets.org/toefl

TOEFL Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Automated Essay Scoring at Scale: A Case Study in Switzerland and Germany

André A. Rupp,[1] Jodi M. Casabianca,[1] Maleika Krüger,[2] Stefan Keller,[2] & Olaf Köller[3]

1 Educational Testing Service, Princeton, NJ
2 Fachhochschule Nordwestschweiz, Olten, Switzerland
3 Leibniz Institute for Science and Mathematics Education, Kiel, Germany

In this research report, we describe the design and empirical findings for a large-scale study of essay writing ability with approximately 2,500 high school students in Germany and Switzerland on the basis of 2 tasks with 2 associated prompts, each from a standardized writing assessment whose scoring involved both human and automated components. For the human scoring aspect, we describe the methodology for training and monitoring human raters as well as for collecting their ratings within a customized platform. For the automated scoring aspect, we describe the methodology for training, evaluating, and selecting appropriate automated scoring models as well as correlational patterns of resulting task scores with scores from secondary measures. Analyses show that the human ratings were highly reliable and that effective prompt-specific automated scoring models could be built with state-of-the-art features and machine learning methods, which resulted in correlational patterns with secondary measures that were in line with general expectations. In closing, we discuss the methodological implications for conducting this kind of work at scale in the future.

Keywords Automated scoring; human scoring; e-rater; *TOEFL*®; Germany; Switzerland; generic scoring model; prompt-specific scoring model; PRMSE; QWK

doi:10.1002/ets2.12249

The use of essay responses in large-scale language assessment has a relatively long history and has spurred various methodological developments regarding task design, scoring approaches, and reporting mechanisms. Despite certain attractive conceptual affordances, one key practical limitation for using essays to assess facets of language proficiency at scale is the associated costs that are incurred in scoring the responses with human raters. As a result, various companies have invested heavily in the development of automated ("machine") scoring technologies for essays that can reduce or, potentially, eliminate the need for human ratings for all — or almost all — responses (for overviews, see, e.g., Shermis & Burstein, 2013; Yan, Rupp, & Foltz, in press).

In this research report, we describe a comprehensive research effort to investigate how automated scoring can be used to evaluate responses by high school learners in Germany and Switzerland to two separate essay writing tasks from the *TOEFL*® test, which is developed in the United States by Educational Testing Service (ETS).[1] This research was part of the Measuring English Writing at Secondary Level (MEWS) study, an international research project funded by the German Research Foundation and the Swiss National Science Foundation. MEWS is the first empirical large-scale study of English writing skills for learners of high school age in Switzerland or Germany (Keller, 2016).

Specifically, we evaluated automated scoring models that were previously created for the international TOEFL learner population, which includes relatively small subsets of German and Swiss learners, as well as various new automated scoring models that were specifically developed for the new sample of learners in this study. We compared the performance of automated scoring systems to that of well-trained human raters who had operational scoring experiences with like prompts. Scientifically speaking, we were interested in exploring the generalizability of human and automated scoring across populations and testing contexts that differed from the original ones for which they were developed.

Doing this work is important for several related reasons. First, it is critical to investigate empirically whether automated scoring models that are developed for a particular population can be directly applied to new populations without much loss in predictive accuracy rather than simply making an untested "plug-and-play" assumption. If existing models can be deployed directly, notable cost savings can be realized relatively quickly, but if new models need to be built, then additional

Corresponding author: A. A. Rupp, E-mail: arupp@ets.org

resource investments need to be made up front as new human ratings are required. Second, in many applied papers on automated scoring, the collection of human scores often gets a relatively short treatment, thereby suggesting that this part of the process is relatively straightforward and requires perhaps few critical design decisions.

However, quite a large number of methodological design decisions have to be made to select, train, calibrate, monitor, and evaluate rater performance so that reliable human ratings are obtained (see Rupp, 2018). Obtaining such ratings is important for at least two related reasons: They serve as the training criterion for new automated scoring models that are being built, and they can be used for score reporting should the automated scoring models perform unsatisfactorily. The latter is important in that the exclusive use of automated scores is generally only advisable in lower stakes assessment contexts or when these models perform exceedingly well; in general, some combination of human and machine scores is preferred for score reporting due to score reliability and construct coverage reasons.

We have organized this research report into four subsequent sections, as follows. In the first section, we provide a methodological overview of the research study context, the tasks and prompts that were used in the study, and the administration design for data collection. In the second section, we discuss the process for collecting human ratings for the resulting essays, which involved the design of a data management system, the design of an administration scheme for administering essays to raters, and various aspects of quality-control monitoring. In the third section, we review our development and evaluation work for the automated scoring models that were preselected and newly built with these human ratings. In the final section, we discuss the various lessons learned from this work for colleagues who may be interested in replicating or adapting it or in designing like efforts.

General Methodology

Writing Task Selection

As noted in the introduction, one of the research goals for this project pertained to empirically investigating whether we could use already developed automated scoring models in this novel-use context without much—or even any—modification. Consequently, it was decided to use the two essay writing tasks that are used on the currently operational TOEFL test developed by ETS; we describe the most important design characteristics of the TOEFL test for the purposes of this research report in the following paragraphs.

One of these tasks is described as the *independent task* and requires the learner to state, explain, and support an opinion based on personal knowledge and experience in an essay of approximately 300 words. The other task is described as the *integrated task* and requires the learner to read a passage of approximately 250–320 words on an academic topic, listen to a lecture of similar length with a conflicting viewpoint on the same topic, and then critically relate the information in the two sources in an essay of approximately 150–225 words in length.

For the purpose of this study, we selected two prompts for each of the two tasks to compare differences in the performance of learners, human raters, and automated scoring models across prompts within the same task type and across the two task types. Learners were randomly assigned to one of the two prompts at each of two administration time points (T1 = September 2016; T2 = June 2017), with the same learners participating at both time points in a repeated-measures design. Because it was not possible to use currently operational prompts for this study due to security concerns for the test overall, we opted to use prompts that had been previously designed and were publicly available on the TOEFL website. Out of the pool of all possible prompts, the research teams in the two countries selected four prompts that had been deemed suitable in terms of content familiarity for the German and Swiss learners. These research teams also tested the suitability of these prompts in small scale pilot studies with select classes to confirm that the tasks and prompts were within the ability range of target student populations and that there were neither notable floor nor ceiling effects.

Specifically, the first independent task, called "Teachers," asked the learners to react to the statement "A teacher's ability to relate well with students is more important than excellent knowledge of the subject being taught." The second independent task, called "Advertising," asked the learners to react to the statement "Television advertising directed toward young children (aged two to five) should not be allowed." The first integrated prompt, called "Voting Machines," included a text and a lecture on the use of different voting systems in the United States, wherein traditional systems were contrasted with computerized systems. The second integrated prompt, called "The Chevalier," included a text and lecture on the memoirs of the Chevalier de Seingalt (1725–1,798), more widely known as Giacomo Casanova. Participants were asked to discuss conflicting statements relating to the reliability of the Chevalier's memoirs.

Figure 1 Screenshots of the platform interface for administering the tasks and recording the written essay responses.

In operational testing practice, TOEFL tasks are administered under strict time constraints of a 20-minute response time for the integrated task and a 30-minute response time for the independent task using a computer platform. The platform interface includes basic writing facilities but does not allow for some other common text-processing activities such as spelling correction, special formatting, or vocabulary look-up through thesauri. For this research study, we designed a customized interface that mimicked the operational administration design for the TOEFL test as closely as possible, with the exception that navigational components and task instructions were translated into German; Figure 1 comprises screenshots of our interface.

In operational testing practice, responses to both tasks are scored using 6-point rubrics, with 0 indicating an empty or off-topic response and scores of 1 – 5 indicating increasing degrees of proficiency. Task scores are then averaged and scaled to a carefully selected reference population using a reporting scale of 0 – 30. When score reports are provided to learners, various interpretational guidelines are provided, which include carefully worded disclaimers that warn learners about common fallacies in misinterpreting scores. Human raters are trained extensively on applying these rubrics consistently, which involves identifying so-called benchmark or anchor responses for each score category and using a specially designed

calibration test to evaluate whether raters are able to perform the task proficiently. We provide more details on the human scoring design in the next section while describing the system that we created for research purposes.

Secondary Measures

To shore up basic validity evidence around the scores for the TOEFL writing tasks, we administered standardized assessments on listening and reading comprehension from the German National Assessment. The tasks from this assessment were designed to monitor the implementation of educational standards in Germany (KMK, 2014), were aligned to the national curricula for the English language classroom, used contextualized scenarios containing authentic language use, and required learners to provide selected responses and short written answers. We estimated expected a posteriori scores for each student using common scaling techniques from item response theory as implemented in Mplus Version 8.0 (Muthén & Muthén, 2017); score reliabilities were above .75 for the two proficiency dimensions. In addition, we collected learners' grades in English, mathematics, and German, along with scores from several background questionnaires; scores from the background questionnaires, however, are not the focus of this study.

Learner Populations

This study involved learners in classes that were almost 2 years before their baccalaureate exams at the first administration (September 2016) and 1 year before their finals at the second administration (June 2017). Note that the educational systems in Germany and Switzerland are generally quite similar in the sense that both countries have "integrated" primary schools (Years 1 – 4 in Germany, Years 1 – 6 in Switzerland) in which all children are schooled together. At the secondary level, both countries have a multi-tiered system that divides students into multiple different tracks.

In Switzerland, learners can get their highest degree (*Matura*) from either an academic track (*Gymnasium*) or a vocational school (*Fachmittelschule*). The Swiss sample used in this study included only learners from the academic track, where selectivity rates in the different regions varied between 15% and 34% of the overall learner population (for more details in German, see, e.g., Schweizerische Koordinationsstelle für Bildungsforschung, 2014). In Germany, there similarly exists an academic track that leads directly to university education (*Gymnasium*) and also various secondary school tracks, including comprehensive schools (*Gesamtschule*) and vocational schools (*berufsorientiertes Gymnasium*), with exact compositions differing by state. In Germany, approximately 40% of an age cohort graduates with the highest degree (*allgemeine Hochschulreife*), and approximately 12% graduate with a subject-specific degree (*fachgebundene Hochschulreife*), the latter allowing them to go to university to study a restricted number of subjects (for more details in German, see, e.g., Autorengruppe Bildungsberichterstattung, 2018).

In our study, every effort was made to create a diverse sample of learners in Switzerland and Germany by selecting schools with differing educational emphases as well as by selecting different cantons (in Switzerland) and by selecting a state in which the average level of English competency is close to the national average in Germany. Data collection took place during regular class sessions in the morning with trained university student assistants and PhD students who supervised the test sessions and instructed all participants. Students took the assessments on a computer and always started the test by writing the two essays from the two prompts, which was followed by the other secondary measures, with a break in between.

In total, we collected 9,628 responses from 2,540 learners across the two administrations, with approximately two thirds of the responses coming from learners in Swiss schools and approximately one third of the responses coming from learners in German schools. The same learners responded to one independent and one integrated prompt at the first administration and the corresponding complementary prompts at the second administration, which resulted in a roughly equal distribution of responses across prompts, as shown in Table 1.

In the following section, we discuss rater selection, rater interface design, and empirical results for the human rating process.

Human Rating Design

Rater Selection

The primary human rating objective was to obtain statistically reliable human ratings that achieved a sufficient degree of construct coverage through the consistent application of the operational TOEFL scoring rubrics. To meet this objective

Table 1 Essay Response Frequencies by Prompt, Country, and Administration

	First administration			Second administration			
Prompt	Switzerland	Germany	Total	Switzerland	Germany	Total	Total
Teachers	847	407	1,254	779	360	1,139	2,393
TV Advertising	906	412	1,318	751	355	1,106	2,424
The Chevalier	872	402	1,274	779	352	1,131	2,405
Voting Machines	882	413	1,295	749	362	1,111	2,406
Total	3,507	1,634	5,141	3,058	1,429	4,487	9,628

Note. As students responded to two prompts at each time point in both countries, the effective number of students in each country is approximately half the total number of responses for each country at each time point.

we decided to use experienced raters from the operational TOEFL pool during a low-volume testing window. Instead of recruiting from the full pool, however, we recruited from a subset of the pool that included raters with historical scoring performance in the upper 30% of the associated rater ability distribution. We eventually hired 36 raters for the first administration and 37 raters for the second administration, with 17 raters rating across both administrations. The raters had an average exact agreement rate of 79% ($SD = 2.4\%$) and 81% ($SD = 4.1\%$) on independent prompt responses and 84% ($SD = 3.4\%$) and 84% ($SD = 2.5\%$) on integrated prompt responses at the first and second administrations, respectively.

Rating Design

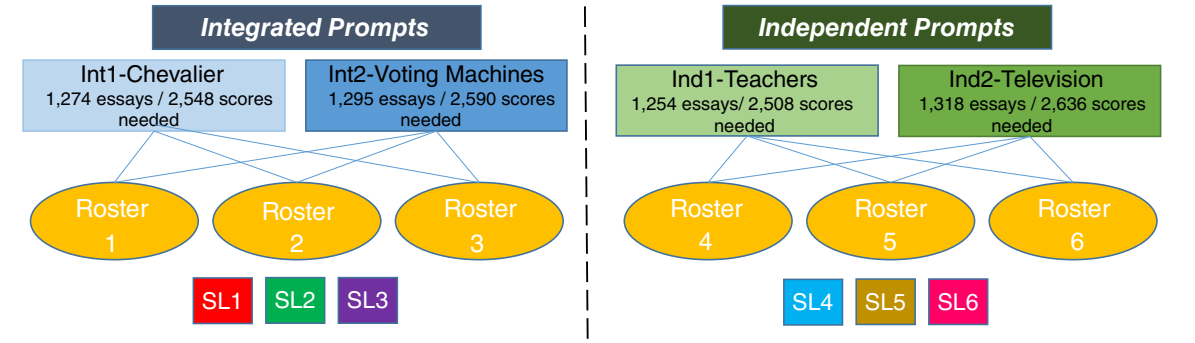
To closely mimic the operational TOEFL scoring process, we also hired *scoring leaders*, who are advanced raters who take on a supervisory role for the raters. They have the ability to read submitted responses, look at as-of-yet unscored responses, review rater performance on validity papers with expert scores, and so on. Each scoring leader manages a roster of raters during scoring to ensure that the raters were correctly applying the scoring guidelines and assigning scores accurately. Typically, scoring leaders manage a roster of about 10 raters; however, in this study, each scoring leader managed a smaller roster of six or nine raters every day during the first and second administration, respectively, to further improve score quality. Each rater was assigned a series of batches of essays to score each day. Specifically, Swiss and German essays were randomly arranged in batches, and batches were randomly assigned to raters to minimize any systematic distribution effects.

All essay responses were double scored, while ensuring that the same rater did not score the same response more than once. Each of the 36 raters was assigned to one of several rosters: six at the first administration and four at the second administration. Moreover, half of the raters were assigned to score independent prompts, and the other half were assigned to score integrated prompts; they had that assignment for the full duration of project scoring, which reduced the time required for the training and calibration that would have been necessary if they had switched between task types. To reduce the risk of score degradation due to undesirable rater effects, however, rosters of raters switched the prompts they scored. To limit training time and leverage scoring experience with a prompt, rosters of raters scored the same prompt for at least 2 days and then switched to a different prompt. To reduce effects of scoring leader assignment, rosters were rotated across scoring leaders within prompt type. Figure 2 gives a pictorial depiction of the rating design for the first administration; the design for the second administration was structurally very similar.

Interface Design

Due to operational demands and certain internal technological constraints, the operational TOEFL scoring platform for the collection of human ratings was not available for use during the time windows required for this study. Note that the scoring platform for raters and scoring leaders is a different platform than the delivery platform for administering the tasks and recording the essay responses for learners. We modified an existing data management system in such a way that it captured the key functionalities of the operational scoring platform; Figure 3 shows screenshots of our scoring platform interface. To assess the impact of the use of our scoring platform instead of the operational one, we conducted a rater and scoring leader survey after the rating effort was completed.

The rating process for essay responses can generally be described in several phases, which include selection, training, calibration, monitoring, remediation, and dismissal. For the purposes of this project, remediation and rater dismissal were not necessary due to the rater selection process. However, training, calibration, and monitoring were still important



Day	Roster 1	Roster 2	Roster 3
Jan 4	Calibration Int1 SL1 Cluster 1 – R1	Calibration Int2 SL2 Cluster 2 – R2	Calibration Int2 SL3 Cluster 2 – R1
Jan 5	Calibration Int1 SL2 Cluster 4 – R1	Calibration Int2 SL3 Cluster 3 – R2	Calibration Int2 SL1 Cluster 3 – R1
Jan 6	Calibration Int1 SL3 Cluster 5 – R1	Calibration Int1 SL1 Cluster 1 – R2	Calibration Int2 SL2 Cluster 6 – R1
Jan 9	Calibration Int2 SL1 Cluster 6 – R2	Calibration Int1 SL2 Cluster 4 – R2	Calibration Int1 SL3 Cluster 5 – R2

Day	Roster 4	Roster 5	Roster 6
Jan 4	Calibration Ind1 SL4 Cluster 7 – R1	Calibration Ind2 SL5 Cluster 8 – R2	Calibration Ind2 SL6 Cluster 8 – R1
Jan 5	Calibration Ind1 SL5 Cluster 10 – R1	Calibration Ind2 SL6 Cluster 9 – R2	Calibration Ind2 SL4 Cluster 9 – R1
Jan 6	Calibration Ind1 SL6 Cluster 11 – R1	Calibration Ind1 SL4 Cluster 7– R2	Calibration Ind2 SL5 Cluster 12 – R1
Jan 9	Calibration Ind2 SL4 Cluster 12 – R2	Calibration Ind1 SL5 Cluster 10 – R2	Calibration Ind1 SL6 Cluster 11 – R2

Figure 2 Human rating design for the first administration.

aspects of the work and required design choices for the interface. For the purpose of rater training, raters were instructed to review a set of materials used in operational rater training. Specifically, they had access to the actual prompt text and media, prompt-specific scoring notes, generic scoring rubrics, and benchmark papers with scores and annotations. Raters were also able to go through a scoring simulation using the benchmark papers, during which they read a benchmark paper and selected the score level from a drop-down box. Upon submitting their score, each rater was provided with instantaneous feedback in the form of an annotation that had been written by a senior scoring leader or test developer, which explained and supported the associated expert score.

For the purpose of rater calibration, each day before starting to score, raters were routed to a web page with training materials and a calibration test. Raters were asked to review the training materials and then to take the calibration test, which consisted of 10 essays. To pass, a rater needed to provide exact scores for seven of the 10 essays. In addition, a rater was not allowed to have more than a two-point difference on any essay (i.e., provide "discrepant" responses). When raters passed the test, they were then routed to a scoring platform so they could proceed with scoring. If they failed, they needed to take a second calibration test. If they failed the second calibration test as well, then they were classified as not qualified to score for that day and were instructed to try again the next day. As expected, however, not many raters failed calibration. At T1, across 4 days of scoring, four raters failed both calibration tests and were excluded from the rating pool on a given day. There were also three instances in which a rater failed the first calibration test and passed the second. At T2, across 2 days of scoring, only two raters failed both calibration tests while five raters failed the first test but passed the second test.

Validity responses, for which consensus scores had been derived from experts during previous operational usage, were also inserted at a rate of approximately 9% per batch, which meant that a typical rater scored roughly seven or eight validity papers per shift, with a read rate of 14 essays per hour and 6 hours of production time. Scoring leaders were able to view the performance on each validity response in a table with hyperlinks to source materials. If an essay received two discrepant scores that were two or more points apart, the essay was flagged for adjudication. Subsequently, the scoring leaders served as adjudicators and assigned a third score. In addition, all responses scored as 0 from one or both raters

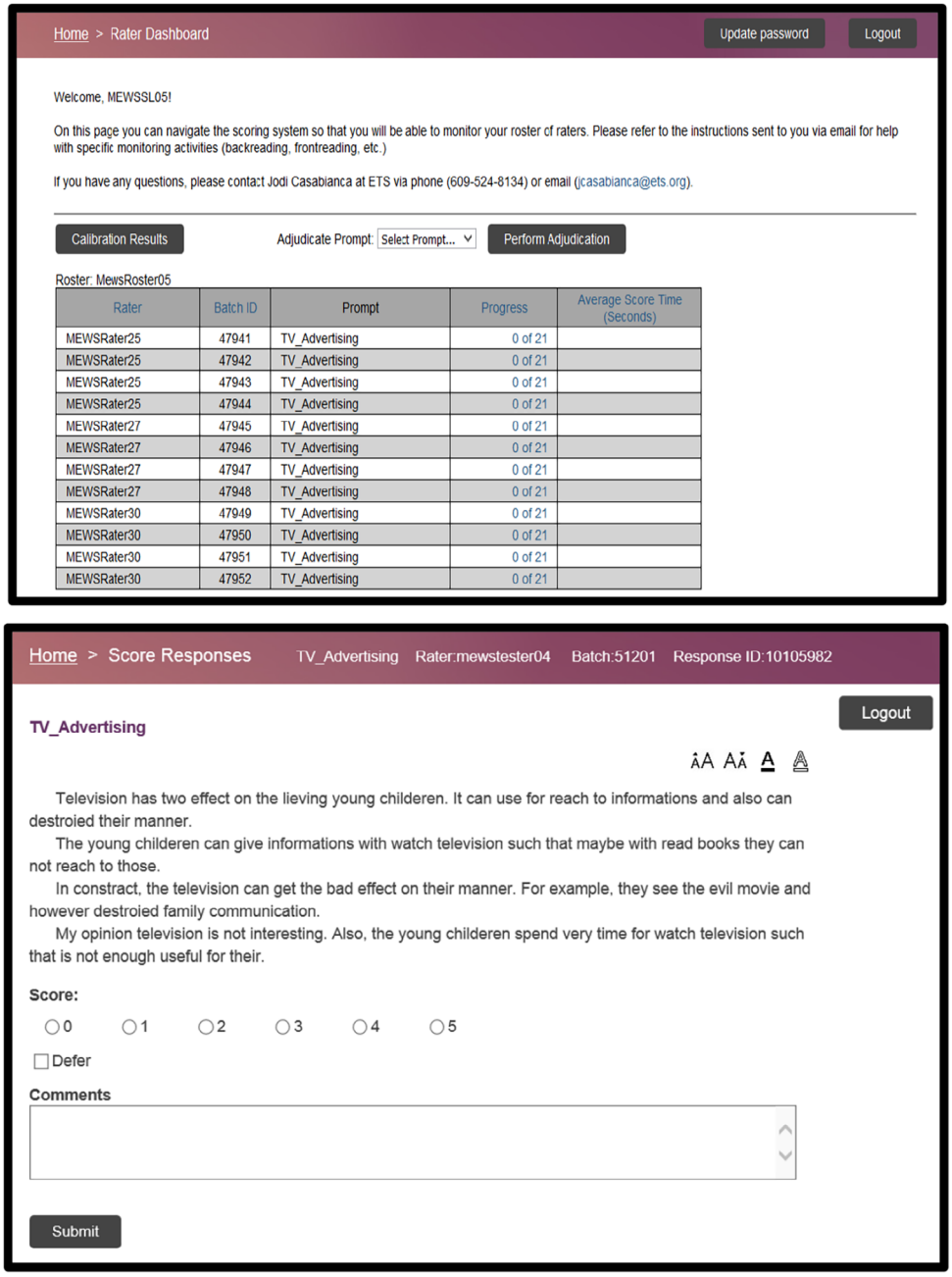


Figure 3 Screenshots of the platform interface for collecting the human ratings from raters(top) and scoring leaders (bottom).

were also pushed to the adjudication queue. For most prompts and administrations, only between 1% and 2% of responses were discrepant and went through the adjudication process with the exception of the Chevalier prompt at T1, which had a 3% adjudication rate.

For the purpose of rater monitoring, the platform was designed to allow for so-called back-reading and front-reading. As the names imply, back-reading refers to scoring leaders reviewing the scores that raters have already assigned to essays to detect any unusual scores, which requires the scoring leaders themselves to assign a score to each response. In contrast, the front-reading functionality allowed them to go into the batches of responses assigned to readers and assign scores to these responses outside of the system, to which they could then compare the scores that raters would give later. Scoring leaders communicated with raters on their performance via e-mail and over the phone. Scoring leaders also wrote end-of-shift reports on raters every day for the benefit of the scoring leader who would be taking over the roster the next day.

Table 2 Human Score Summary Statistics Based on Task Score Across Administrations

		Switzerland				Germany				Total			
Time	Prompt	*N*	Mean	Median	*SD*	*N*	Mean	Median	*SD*	*N*	Mean	Median	*SD*
Independent													
T1	Teachers	847	3.38	3.50	0.68	407	3.20	3.00	0.71	1,254	3.32	3.50	0.70
T2	Teachers	779	3.43	3.50	0.73	360	3.27	3.00	0.77	1,139	3.38	3.50	0.75
T1 + T2	Teachers	1,626	3.40	3.50	0.71	767	3.24	3.00	0.74	2,393	3.35	3.50	0.72
T1	TV advertising	906	3.22	3.00	0.74	412	3.07	3.00	0.72	1,318	3.17	3.00	0.74
T2	TV advertising	751	3.19	3.00	0.76	355	3.04	3.00	0.76	1,106	3.14	3.00	0.76
T1 + T2	TV advertising	1,657	3.20	3.00	0.75	767	3.06	3.00	0.74	2,424	3.16	3.00	0.75
Integrated													
T1	The Chevalier	872	2.74	3.00	1.05	402	2.17	2.00	1.10	1,274	2.56	3.00	1.10
T2	The Chevalier	779	3.04	3.00	1.09	352	2.50	2.75	1.09	1,131	2.87	3.00	1.11
T1 + T2	The Chevalier	1,651	2.88	3.00	1.08	754	2.33	2.50	1.11	2,405	2.71	3.00	1.11
T1	Voting machines	882	2.89	3.00	0.80	412	2.63	3.00	0.90	1,294	2.81	3.00	0.84
T2	Voting machines	749	3.10	3.00	0.83	362	2.96	3.00	0.84	1,111	3.06	3.00	0.84
T1 + T2	Voting machines	1,631	2.99	3.00	0.82	774	2.79	3.00	0.89	2,405	2.92	3.00	0.85

Note. Since each response was targeted for double human scoring but operational adjudication procedures were used, the effective number of human ratings reflected in these statistics is approximately twice the number of responses/learners (*N*). T1 = time/Administration 1; T2 = time/Administration 1; T1 + T2 = both administrations combined.

Empirical Findings

Table 2 shows the resulting human score summary statistics based on the task-level score assigned to each response, which was a combination of the two human scores (explained in detail in the "Model Building" section). The data in Table 2 show, for the combined German and Swiss data across both T1 and T2, that the mean scores for the two independent prompts ($M = 3.35$; $M = 3.16$) were higher than those of the integrated prompts ($M = 2.71$; $M = 2.92$) and that learners received more similar scores for the former ($SD = .72$; $SD = .75$) than for the latter ($SD = 1.11$; $SD = .85$).

They also show that the Swiss learners outperformed the German learners as reflected in mean scores that were up to about a half point higher in the former sample. The score distributions were overall similar at the two administrations for the independent prompts, but mean scores increased slightly for the integrated prompts at the second administration. This was probably due to a mixture of task familiarity effects, additional instructional exposure to this task type between administrations, and slight learning gains relative to this task type; exact reasons cannot be discerned without proper experimental designs, of course.

Table 3 shows interrater agreement statistics by prompt based on the original two scores provided by the raters (i.e., without considering adjudication) for both administrations separately (T1, T2) as well as pooled $(T1 + T2)^2$; the upper half of the table provides statistics for the independent prompts, and the bottom half provides statistics for the integrated prompts. The data in Table 3 show moderate to high agreement among pairs of human raters, even though the agreement was lower for the independent task (e.g., quadratic-weighted kappa [QWK] = .670; QWK = .639, for the pooled data) than the integrated task (e.g., QWK = .865; QWK = .775, for the pooled data). In most cases, the agreement was very similar across the German and Swiss samples and decreased only slightly across the two administrations in a few cases.

The one notable exception was the TV Advertising prompt at T2, especially in the German sample (e.g., QWK = .568 for the German sample; QWK = .646 for the Swiss sample). After a review of a random selection of essays that received conflicting scores, we found that some learners were confused with the prompt and wrote responses focused on children and television but did not incorporate ideas around television advertising. It is unknown if this happened at a higher rate in the German sample more generally; however, at T2, the German sample did exhibit lower levels of motivation based on other information we collected. It is likely that both of these issues led to borderline off-topic responses, which are relatively difficult to score consistently, thereby yielding lower human–human agreement. Despite this pattern for TV Advertising at T2, the human ratings were overall sufficiently consistent across prompts and administrations so that they could be used with reasonable confidence for score reporting, if desired, as well as for the development of automated scoring models, which we discuss in the next section.

Table 3 Human – Human Agreement Statistics Across Administrations

		Switzerland				Germany				Total			
Time	Prompt	% Exact	% Adj	QWK	*r*	% Exact	% Adj	QWK	r	% Exact	% Adj	QWK	*r*
Independent													
T1	Teachers	64.5	99.5	.669	.670	64.9	99.8	.697	.698	64.6	99.6	.682	.683
T2	Teachers	59.2	98.8	.648	.648	62.1	99.2	.693	.697	60.1	98.9	.665	.666
T1 + T2	Teachers	62.0	99.2	.656	.656	63.5	99.5	.689	.690	62.5	99.3	.670	.670
T1	TV Advertising	60.4	98.6	.665	.665	64.1	98.8	.669	.670	61.5	98.6	.669	.669
T2	TV Advertising	57.8	98.4	.646	.649	51.6	96.9	.568	.573	55.8	97.9	.623	.626
T1 + T2	TV Advertising	59.1	98.5	.646	.647	58.4	97.9	.616	.618	58.9	98.3	.639	.640
Integrated													
T1	The Chevalier	68.8	99.1	.849	.856	73.0	98.5	.869	.881	70.1	98.9	.863	.871
T2	The Chevalier	67.0	98.6	.848	.848	69.9	100.0	.873	.875	67.9	99.0	.862	.863
T1 + T2	The Chevalier	68.0	98.8	.851	.853	71.6	99.2	.874	.876	69.1	99.0	.865	.867
T1	Voting Machines	67.5	99.3	.763	.764	67.0	99.3	.804	.805	67.3	99.3	.782	.783
T2	Voting Machines	68.9	98.7	.756	.758	67.8	98.9	.758	.768	68.5	98.7	.758	.762
T1 + T2	Voting Machines	68.2	99.0	.763	.765	67.3	99.1	.790	.792	67.9	99.0	.775	.777

Note. QWK = quadratic-weighted kappa.

Automated Scoring Design

The term *automated scoring* subsumes a wide variety of methodological approaches for model building and evaluation, especially once various task formats (e.g., short answer, extended response, interactive performance) across various modalities (e.g., speaking, writing, multimodal output) are considered. In the following, we describe only the key steps for the kind of prototypical model that is common in applications of automated essay scoring.

Feature Space

A first step in the development process for an automated scoring model for essays is processing the digitally collected written responses via computational routines, which results in a set of statistical variables — called *features* — that can then be used as predictor variables in statistical models to yield predicted human scores for these essays. The development of features is scientifically grounded in the disciplines of natural language processing, computational linguistics, and computer science, to name a few key disciplines, with boundaries between these disciplines being somewhat fuzzy (see, e.g., Deane, 2006; Jurafsky & Martin, 2009; Quinlan, Higgins, & Wolff, 2009).

At ETS, features go through a rigorous process of internal vetting, which includes an interdisciplinary review along a variety of conceptual and empirical dimensions that pertain to feature definition, computational operationalization, statistical performance, and processing speed, before they are deployed operationally. The major operational automated scoring engine for essay responses that uses these features at ETS is referred to as the *e-rater*® automated scoring engine, which undergoes periodic updates for improvement and general maintenance (e.g., Burstein, Tetreault, & Madnani, 2013). In this research project, we used both the currently operational engine version and a research version of the engine from a different server architecture into which two new preoperational features, *discourse* and *source-use*, had been embedded.

We further differentiate between features at two different levels of computational and semantic grainsize — *microfeatures* and *macrofeatures* — with macrofeatures typically created as an aggregation of microfeatures. In addition, we differentiate between features that are useful for building *generic scoring models* (i.e., models that can be used for collections of prompts) and features for *prompt-specific scoring models* (i.e., models that are built for specific prompts). The latter set of features typically makes use of the vocabulary information from the content that is unique to each prompt, which is why they do not generalize across prompts.

Figure 4 provides a graphical overview of the features that we used in this research. In particular, macrofeatures are shown at the top and included grammar, usage, mechanics, organization, development, discourse, collocations and prepositions, average word length, median word frequency, and sentence variety. Two macrofeatures measured the use of vocabulary and were used only in prompt-specific models, as shown in the bottom-right box. Finally, two separately trained versions of the source-use feature were used, but only for the integrated task, which is why it is shown in two versions: once on the top for generic models and once in the box for prompt-specific models. Appendix A contains a

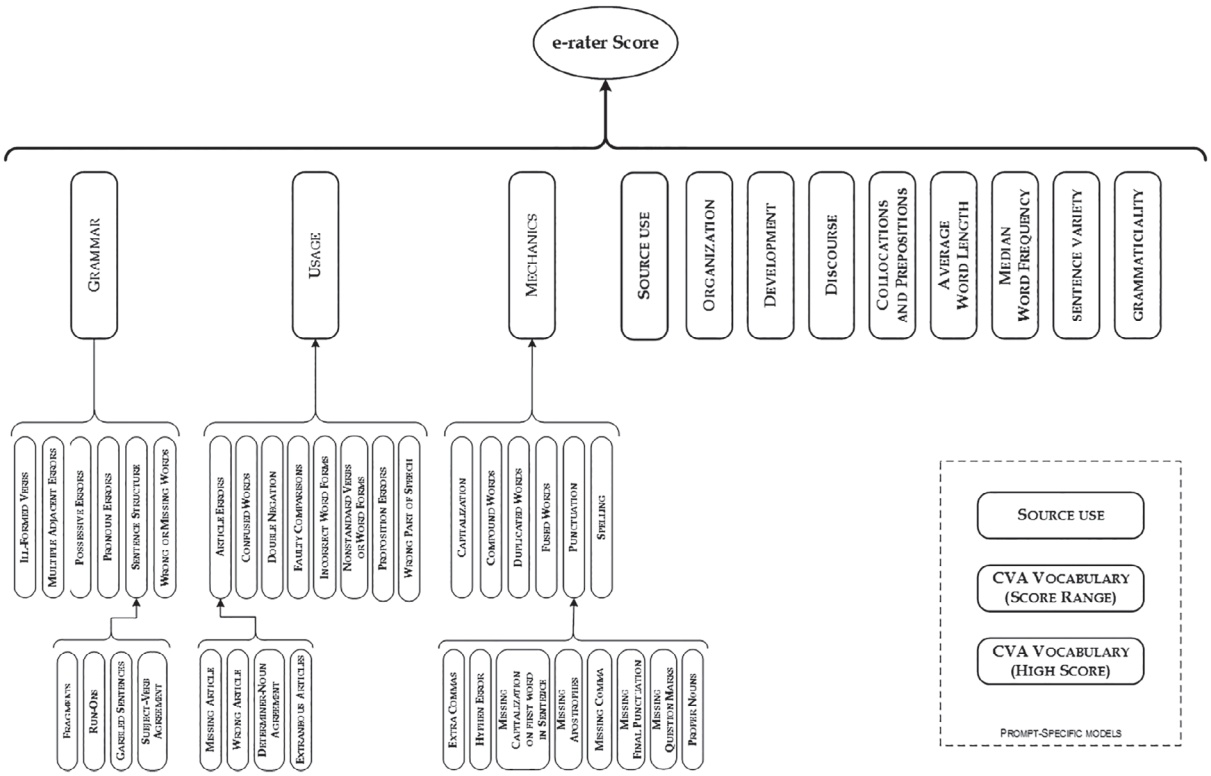


Figure 4 Construct representation in the e-rater engine via microfeatures and macrofeatures. CVA = content vector analysis.

brief description of each of the macrofeatures, based, in large part, on Attali and Burstein (2006), while Figure 4 provides an overview of which features were included in which models.

Case Removal

During the process of feature extraction, the *e-rater* engine computes a set of so-called advisory flags to identify essays that exhibit undesirable properties that lead to untrustworthy scores. Specifically, we excluded independent and integrated essays that were considered too brief, had an excessive length, or had an excessive number of errors via internal flagging thresholds that had been set based on previous operational data. We also excluded independent and integrated essays that were written by foreign exchange students whose first language was English, contained evidence of personal emotional content, contained notable amounts of text deletion, or were either abnormally or deliberately terminated. Finally, essays that were flagged by human raters as unusual were similarly excluded. Common reasons for case exclusion by human raters included situations in which learners commented that they were unable to hear audio (for the integrated task only) or reported other technical difficulties (for both tasks). The total number of such cases ranged from less than 2% to no more than 6% across prompts and administrations.

Model Building

The models we report on for the purpose of this research report were built using the combined data across the German and Swiss populations to ensure that the resulting scoring model could be applied across countries (i.e., without yielding country-specific influences of scoring features on predicted scores). After we removed cases with advisory flags, we randomly split the data into two nonoverlapping subsets of equal size, a *model building partition* and a *model evaluation partition*; models were trained on the model-building partition and evaluated on the model evaluation partition. The outcome variable for training all models was either (a) the average of the first and second human ratings (if no adjudication was needed), (b) the average of all three ratings/the adjudicator rating (in cases of patterns like 1 – 2 – 3, where 2 was the adjudicator), or (c) the average of the two closest human ratings (in cases of patterns like 1 – 3 – 4).

We built models in several different ways, which included building generic and prompt-specific models using multiple linear regression as well as five additional machine learning techniques. In all multiple linear regression models, any feature variable that had a negative weight was iteratively excluded from the model in a process that mimicked strict nonnegative least squares estimation. This process is used for building operational multiple linear regression models to preserve construct representation via an interpretable positive weighting scheme for features based on the way they are scaled. In terms of machine learning techniques, we used (a) ridge regression, (b) lasso regression, (c) elastic nets, (d) support vector regression, and (e) linear support vector regression (for an overview, see, e.g., Dangeti, 2017).

In total, we built 96 new models with operational and experimental features. Specifically, with the data from each administration, we built 36 models: a total of 12 generic models (2 tasks × 6 estimation methods), all without prompt-specific features but with experimental features, as well as 24 prompt-specific models (4 prompts × 6 estimation methods), all with prompt-specific and experimental features; this resulted in 72 models across the two administrations. We then built an additional 24 prompt-specific models in the same vein using the pooled data from both administrations; we did not rebuild generic models because they had underperformed at administrations in preliminary evaluations. Table 4 provides an overview of all the different models and the features they contained. All models were trained using RSMTool Version 5.6 (Madnani, Loukina, von Davier, Burstein, & Cahill, 2017) using default parameters with grid search, where applicable.

Model Evaluation

The statistical performance of the scoring models was evaluated using univariate and bivariate statistics as well as associated plots of distributions. Specifically, we inspected core parametric moments for score distributions (e.g., mean, standard deviation, kurtosis, and skewness), quantiles (e.g., median, interquartile range), exact and adjacent agreement rates (percentage [adjacent] agree), QWK, and Pearson correlation coefficients (r), as well as standardized mean differences (SMD) between average human and machine scores (i.e., Cohen's d).

Finally, we computed mean squared error (MSE), which is the average of the squared differences between an estimator (e.g., machine score) and the quantity to be estimated (e.g., human score), as well as true-score prediction statistics based on the proportional reduction in mean squared error (PRMSE) statistic (see, e.g., Haberman, 1996, p. 280). It can be shown that the MSE_H/PRMSE_H statistics inform how well human true scores can be predicted from human observed scores, while MSE_M/PRMSE_M statistics inform how well human true scores can be predicted from machine scores; equations for the variants of MSE and PRMSE we used in this study are provided in Appendix B.

All statistics were computed using so-called bounded e-rater scores (i.e., real-valued numbers ranging from 0.5002 to 5.4998), while the percentage [adjacent] agree statistics were computed using rounded e-rater scores (i.e., integer scores); the boundary values for the bounded scores were chosen to eliminate the influence of a few cases with extreme values, as those cases were recoded to take on the boundary values. For more information on the computational details and distributional properties of these statistics, please refer to a standard introductory book on statistics (e.g., Lomax & Hahs-Vaughn, 2012) or an introductory book on categorical data analysis (e.g., Agresti, 2013) as well as the references cited therein. For more information on the use of these statistics for general quality-control evaluations, please refer to Williamson, Xi, and Breyer (2012) and references therein.

Covariate information on learners was used solely for the purpose of defining relevant population subgroups and inspecting key performance statistics — most importantly, QWK, SMD, MSE, and PRMSE — as well as quality-control checks on the random sample splitting procedure to create the two data partitions. Subgroups with very small sample sizes were not used to inform model performance recommendations or decisions — we compared performance only using the variables countries (Germany and Switzerland) and gender (male, female, unknown).

Empirical Findings

Before reporting on the performance of the prompt-specific models, we first briefly review the performance of the two currently operational generic models using our data from the first administration only. Those models are listed as models M1a and M1b in Table 4 and had previously been built using a large sample of the TOEFL population. We do this only to demonstrate the differences between using an off-the-shelf scoring model and models that are retrained for a new sample while staying within the same model family. We refer to models that we built for our sample as "customized" generic and prompt-specific models, which are listed as models M2a – M2b and M3a – M3d in Table 4, respectively.

Table 4 Summary of Models and Feature Sets (T1, T2)

Model type	Model ID	Task/prompt	Operational engine version	Research engine version	Experimental features		Prompt-specific features		Total features in model	Data use
					Source-use	Discourse	Vocabulary score range	Vocabulary high score		
Generic (off-the-shelf)	M1a	Independent	X						9	T1
	M1b	Integrated	X						9	T1
Generic (customized)	M2a	Independent		X		X			10	T1, T2
	M2b	Integrated		X	X (V1)	X			11	T1, T2
Prompt-specific (customized)	M3a	Independent (teachers)		X		X	X	X	12	T1, T2, (T1 + T2)
	M3b	Independent (TV advertising)		X		X	X	X	12	T1, T2, (T1 + T2)
	M3c	Integrated (The Chevalier)		X	X (V2)	X	X	X	13	T1, T2, (T1 + T2)
	M3d	Integrated (voting machines)		X	X (V2)	X	X	X	13	T1, T2, (T1 + T2)

Note. V1 = version 1 of source-use feature for generic models; V2 = version 2 of source-use feature for prompt-specific models (see Figure 4); X = feature included in scoring model.

Performance for Generic Models (T1)

Table 5 provides a series of human – machine agreement statistics based on the model evaluation partition from the first administration; the first two rows provide statistics for the independent task, and the bottom two rows provide statistics for the integrated task.

The customized models showed an improvement over the off-the-shelf model; this effect was similarly pronounced for the independent task ($r = .730$ vs. $r = .769$ for the off-the-shelf and customized models, respectively) and the integrated task ($r = .590$ vs. $r = .613$ for the off-the-shelf and customized models, respectively). The agreement statistics were larger for the German sample, in particular for the integrated task, for which correlations were more than .07 higher in the German sample compared to the Swiss sample. The agreement statistics were notably larger for the independent task than for the integrated task, which was expected due to the more complete construct coverage for the former despite the inclusion of the two experimental features in the models for the latter. The generic models for the integrated task showed a level of performance that would generally be considered unacceptable for reporting if used as sole scores and in high-stakes contexts.

A similar performance pattern was observed via the SMD statistic, which showed smaller absolute differences for the independent task (SMD = .044 vs. SMD = −.005 for the off-the-shelf and customized models, respectively) than for the integrated task (SMD = .608 vs. SMD = −.018 for the off-the-shelf and customized models, respectively). These differences are due to the fact that the intercept was not adjusted in the off-the-shelf application; if such an adjustment were made, then these effects would be reduced. For three of the four models, the SMD statistics were also notably larger for the German sample, which was likely because the overall sample included about twice as many Swiss learners.

Table 6 provides a series of statistics that describe the associated precision of the score estimation for human true scores (MSE_H, $PRMSE_H$) and machine scores (MSE_M, $PRMSE_M$).

As before with the agreement statistics, the MSE_M values improved for the customized models relative to the off-the-shelf generic model for both the independent task ($MSE_M = 0.276$ vs. $MSE_M = 0.199$ for the off-the-shelf and customized models, respectively) and the integrated task ($MSE_M = 1.149$ vs. $MSE_M = 0.535$ for the off-the-shelf and customized models, respectively). Moreover, the MSE_M values were lower than the corresponding MSE_H values for the independent task but were higher for the integrated task, which was again a reflection of the more incomplete construct coverage for the integrated task.

Performance of Customized Prompt-Specific Models (T1 + T2)

For the purposes of this research report, we do not delve into the performance of the off-the-shelf generic models further and, instead, focus on building the best-performing customized models. These models were prompt specific, included the two experimental features, were built on pooled data, and utilized a support vector machine estimation approach (e.g., Drucker, Burges, Kaufman, Smola, & Vapnik, 1997). Detailed results for these models are shown in Tables 7 – 9, broken down by country, in alignment with the presentation of information in previous tables. That is, Table 7 shows the predicted score distribution statistics, Table 8 shows the agreement statistics, and Table 9 shows the score precision statistics.

As seen in Table 7, for the overall sample, mean machine scores were higher for the independent prompts ($M = 3.37$; $M = 3.18$) than for the integrated prompts ($M = 2.80$; $M = 2.98$), with more consistent scores for the former prompts ($SD = 0.62$; $SD = 0.72$) than for the latter prompts ($SD = 1.03$; $SD = 0.80$). This is consistent with the human score distributions shown in Table 2 as the machine score means are generally within 0.05 of the human score means, but with standard deviations of machine scores that are between 0.03 and 0.10 lower than the human score standard deviations. Again, just as with human scores, the Swiss learners are shown to outperform the German learners for all four prompts.

As seen in Table 8, human – machine agreement was satisfactorily high for all prompt-specific models, as correlations ranged from $r = .761$ to $r = .809$ for the independent prompts and from $r = .698$ to $r = .825$ for the integrated prompts in the two samples. As a reference point, consider the results from the generic models at T1 in Table 5, where human – machine agreement for customized models ranged from $r = .769$ to $r = .771$ for the independent prompts and from $r = .578$ to $r = .650$ for the integrated prompts. In terms of mean score agreement, SMD values for the best-performing models were almost always less than .10 and typically less than .05 in the two samples, which was acceptable and a notable improvement over the SMD values for the generic models in Table 5.

Table 5 Human–Machine Agreement Statistics for Generic Models Based on Data From First Administration (T1)

	Switzerland						Germany						Total					
Model	*N*	% Exact	% Adj	QWK	*r*	SMD	*N*	% Exact	% Adj	QWK	*r*	SMD	*N*	% Exact	% Adj	QWK	*r*	SMD
Independent																		
Off-the-shelf	906	37.6	98.9	0.725	0.725	0.017	461	33.6	98.5	0.734	0.741	0.128	1,367	36.3	98.8	0.729	0.730	0.044
Customized	890	42.5	99.4	0.740	0.770	−0.068	458	42.6	98.7	0.754	0.771	0.107	1,348	42.5	99.2	0.746	0.769	−0.005
Integrated																		
Off-the-shelf	926	17.2	76.5	0.473	0.549	0.548	424	13.2	73.8	0.549	0.636	0.535	1,350	15.9	75.6	0.511	0.590	0.608
Customized	903	29.2	90.7	0.521	0.578	−0.075	382	26.2	89.8	0.603	0.650	0.102	1,285	28.3	90.4	0.560	0.613	−0.018

Note. The sample sizes reported in this table reflect the number of learners (*N*) who were included in the model evaluation process for each modeling approach and task combination. Slight discrepancies in *N* across modeling approaches reflect differences in the advisory flags and exclusions for different scoring engines. For example, 906 Swiss learners were in the model evaluation sample for the off-the-shelf model for the Independent task. Similarly, 926 Swiss learners were in the model evaluation sample for the off-the-shelf model for the Integrated task, with the majority of these cases overlapping. Adj = adjusted; QWK = quadratic-weighted kappa; SMD = standardized mean differences.

Table 6 Score Precision Statistics for Generic Models Based on Data From First Administration (T1)

	Switzerland					Germany					Total				
Model	*N*	MSE_H	MSE_M	$PRMSE_H$	$PRMSE_M$	*N*	MSE_H	MSE_M	$PRMSE_H$	$PRMSE_M$	*N*	MSE_H	MSE_M	$PRMSE_H$	$PRMSE_M$
Independent															
Off-the-shelf	906	0.421	0.272	0.788	0.667	461	0.399	0.284	0.802	0.685	1,367	0.413	0.276	0.794	0.672
Customized	890	0.424	0.192	0.774	0.766	458	0.397	0.211	0.804	0.740	1,348	0.415	0.199	0.786	0.751
Integrated															
Off-the-shelf	926	0.333	1.130	0.897	0.337	424	0.368	1.191	0.905	0.446	1,350	0.344	1.149	0.903	0.386
Customized	903	0.337	0.526	0.893	0.375	382	0.395	0.558	0.896	0.471	1,285	0.354	0.535	0.897	0.419

Note. The sample sizes reported in this table reflect the number of learners (*N*) who were included in the model evaluation process for each modeling approach and task combination. Slight discrepancies in *N* across modeling approaches reflect differences in the advisory flags and exclusions for different scoring engines. For example, 906 Swiss learners were in the model evaluation sample for the off-the-shelf model for the Independent task. Similarly, 926 Swiss learners were in the model evaluation sample for the off-the-shelf model for the Integrated task, with the majority of these cases overlapping. *H* = human; *M* = machine; MSE = mean squared error; PRMSE = proportional reduction in mean squared error.

Table 7 Machine Score Summary Statistics for Best-Performing Models Based on Pooled Data Across Administrations (T1 + T2)

	Switzerland					Germany					Total				
Prompt	*N*	Mean	*SD*	Median	IQR	*N*	Mean	*SD*	Median	IQR	*N*	Mean	*SD*	Median	IQR
Teachers	792	3.41	0.60	3.50	1	380	3.28	0.65	3.50	0.5	1,172	3.37	0.62	3.50	1
TV Advertising	797	3.22	0.70	3.00	0.5	398	3.10	0.74	3.00	1	1,195	3.18	0.72	3.00	1
The Chevalier	773	2.97	0.96	3.00	1	353	2.43	1.09	2.50	2	1,126	2.80	1.03	3.00	1.5
Voting Machines	801	3.06	0.78	3.00	1	355	2.79	0.81	2.50	1	1,156	2.98	0.80	3.00	1

Note. The sample sizes reported in this table reflect the number of learners (*N*) in the model evaluation sample with machine scores estimated from the best-fitting automated scoring model based on the pooled data (T1 + T2). IQR = Inter-quartile range.

Table 8 Human – Machine Agreement Statistics for Best-Performing Models Based on Pooled Data Across Administrations (T1 + T2)

	Switzerland						Germany						Total					
Prompt	*N*	% Exact	% Adj	QWK	*r*	*d*	*N*	% Exact	% Adj	QWK	*r*	*d*	*N*	% Exact	% Adj	QWK	*r*	*d*
Teachers	792	46.1	99.0	0.763	0.772	0.007	380	39.5	99.5	0.755	0.761	0.068	1,172	43.9	99.1	0.762	0.770	0.021
TV Advertising	797	41.3	99.4	0.805	0.806	0.017	398	43.0	99.2	0.808	0.809	0.034	1,195	41.8	99.3	0.807	0.808	0.020
The Chevalier	773	30.0	95.5	0.792	0.795	0.028	353	31.7	95.8	0.825	0.825	0.015	1,126	30.6	95.6	0.814	0.815	0.029
Voting Machines	801	34.8	96.0	0.697	0.698	0.057	355	34.1	96.3	0.741	0.747	−0.107	1,156	34.6	96.1	0.715	0.715	0.005

Note. The sample sizes reported in this table reflect the number of learners (*N*) in the model evaluation sample with machine scores estimated from the best-fitting automated scoring model based on the pooled data (T1 + T2). QWK = quadratic-weighted kappa; Adj = adjusted.

As seen in Table 9, the MSE_M values were again much smaller for the independent task, for which they ranged from 0.192 to 0.218 for the two samples, than for the integrated task, for which they ranged from 0.341 to 0.409 for the two samples. Moreover, for the two independent prompts, the MSE_M values were generally less than half the size of the MSE_H values, while for the two integrated prompts, they were much more similar to one another, with the MSE_M values slightly exceeding the MSE_H values. These patterns are in line with expectations because, as we saw before in Tables 3, 5, and 8, there is generally higher agreement between human raters on integrated tasks and higher human – machine agreement on independent tasks due to the more limited but consistent construct coverage of the automated scoring model for the latter.

The PRMSE statistics reflect these differences as well, even though the differences in magnitudes are not as large for the independent task. Similar to other effect size metrics, there is no singular PRMSE threshold that is used across all contexts to make a judgment about score quality. For example, using a pretty conservative value of .90, all of the scoring models would show insufficient predictive accuracy to support a sole score use, but if one were willing to accept more lenient values of approximately .80 or even .75, then one might consider reporting these scores, at least for the independent tasks.

Finally, Table 10 shows the correlations of combined human and machine scores (i.e., the average of all available human ratings plus the machine score from the best-performing model) with scores from the standardized national assessments of reading and listening comprehension as well as school grades in English, German, and mathematics.

Table 9 Score Precision Statistics for Best-Performing Models Based on Pooled Data Across Administrations (T1 + T2)

	Switzerland					Germany					Total				
Prompt	N	MSE_H	MSE_M	$PRMSE_H$	$PRMSE_M$	N	MSE_H	MSE_M	$PRMSE_H$	$PRMSE_M$	N	MSE_H	MSE_M	$PRMSE_H$	$PRMSE_M$
Teachers	792	0.405	0.193	0.784	0.760	380	0.382	0.218	0.806	0.719	1,172	0.398	0.201	0.794	0.747
TV Advertising	797	0.476	0.192	0.769	0.845	398	0.495	0.202	0.769	0.851	1,195	0.482	0.195	0.770	0.848
The Chevalier	773	0.344	0.409	0.919	0.688	353	0.365	0.406	0.920	0.740	1,126	0.351	0.408	0.923	0.719
Voting Machines	801	0.355	0.360	0.853	0.571	355	0.338	0.341	0.879	0.635	1,156	0.349	0.354	0.862	0.594

Note. The sample sizes reported in this table reflect the number of learners (N) in the model evaluation sample with machine scores estimated from the best-fitting automated scoring model based on the pooled data (T1 + T2). H = human; M = machine; MSE = mean squared error; PRMSE = proportional reduction in mean squared error.

Table 10 Correlations Between Scores on Essays and Secondary Measures (T1, T2)

	Prompt			
Assessment	Teachers	TV Advertising	The Chevalier	Voting Machines
Switzerland				
Listening comprehension	.554/.576	.591/.563	.621/.694	.568/.640
Reading comprehension	.466/.504	.476/.474	.510/.559	.474/.537
English grade	.465/.458	.477/.436	.484/.373	.423/.454
German grade	.226/.273	.242/.221	.202/.113	.181/.293
Math grade	$-.012^{ns}$/.085	$.068^{ns}$/$.029^{ns}$	$.040^{ns}$/.081	.154/.115
Germany				
Listening comprehension	.578/.594	.604/.591	.614/.725	.616/.682
Reading comprehension	.463/.486	.536/.522	.462/.592	.415/.578
English grade	.557/.516	.481/.532	.489/.486	.428/.436
German grade	.325/.400	.332/.337	.320/.273	.286/.344
Math grade	.193/.227	.144/.212	.217/.222	.201/.293

Note. ns = nonsignificant coefficients, $p > .05$. All other coefficients are significant, $p < .05$. Coefficient structure is T1/T2.

Correlational patterns were in line with expectations, especially within the German and Swiss samples; we will briefly summarize the pattern across samples here for simplicity. The scores from all four essay prompts showed substantial positive correlations with the scores from the listening and reading comprehension assessments, which ranged from $r = .554$ to $r = .725$ and from $r = .415$ to $r = .592$, respectively. Similarly, the highest correlations of all four essay scores across samples were found with the school grade in English, which ranged from $r = .373$ to $r = .557$. Lower correlations were found for the school grade in German, which ranged from $r = .113$ to $r = .400$, and the smallest correlations were found with the school grade in mathematics, which ranged from $r = -.012$ to $r = .293$.

Discussion

Put simply, our results show that well-trained human raters were able to score our learner responses reliably for all four prompts and that custom-built prompt-specific automated scoring models created with modern computational tools performed generally satisfactorily for our mixed population and tasks. The findings in this research report underscore how the use of automated scoring technology in a large-scale assessment context typically requires new research investments rather than a simple plug-and-play approach with previously developed systems or models.

The degree to which substantial investments are necessary depends on a variety of factors. Generally speaking, the more operational deployment restrictions are put in place (e.g., requiring comparable scoring models across prompts, requiring strong validity evidence, requiring robust detection of aberrant responses), the more additional research is likely to be necessary. If mature scoring systems are in place and effective scoring models have already been built based on prior work, then it may also be possible to adjust those models using simple mechanisms (e.g., mean/intercept shifts), although human ratings are still needed to evaluate the performance of the adjusted models. In general, industry specialists can attest that starting an automated scoring enterprise "from scratch" is a decidedly nontrivial endeavor, no matter what the use context is, and requires various computational, scientific, and managerial investments. As a

result, it is sometimes easier to purchase scoring services from a trustworthy vendor rather than building systems anew, unless a long-term development horizon with a clear vision for assessment and learning can help lay out a road map.

In the following sections, we briefly discuss a few strengths and limitations of this study as well as lessons learned from this work that might be helpful to consider for colleagues who would like to pursue similar scoring endeavors.

Strengths

One of the greatest strengths of this study was the care with which human ratings were collected. The practices implemented here mimicked those of operational work in a conservative high-stakes context and, thus, led to high interrater agreement within a relatively efficient rating process. This allowed for the building of customized automated scoring models and avoided the garbage-in-garbage-out phenomenon that is at play whenever relatively noisy human ratings are used to train such models. Moreover, having such ratings available allows for the direct reporting of such scores to stakeholders, which typically has a higher acceptance than reporting a mixture of human and machine scores or even machine scores by themselves (see, e.g., Wood, 2017).

An additional strength was the use of a relatively mature automated scoring system for this research, which in this case was the e-rater scoring engine with its associated sophisticated feature set and robust computational architecture. While more simplistic automated scoring approaches can sometimes yield surprisingly good results, they nevertheless reach performance ceilings (e.g., Lottridge, 2018). These are due to a variety of factors that include a lack of sophisticated disambiguation routines for features that impact classification accuracy (e.g., differentiating accurately between different possible grammatical error types with the aid of contextual factors) and an associated lack of more sophisticated computational features that cover more conceptually complex construct aspects (e.g., maturity of reasoning, coherence of discourse). Whenever scores need to carry complex meaning, no matter what the exact stakes of the assessment, these limitations can become troublesome.

Finally, some secondary methodological design decisions in this study were made carefully as well. The research platform interface mimicked the operational platform interface, even though it was not exactly identical to it. We also collected basic validity evidence to learn about the trustworthiness of the resulting classifications. Finally, even though it is not reported herein, we also carefully designed and implemented a standard-setting procedure to align reported scores with the Common European Framework of Reference for Languages (Fleckenstein, Keller, Krüger, Tannenbaum, & Köller, 2019).

Limitations

One of the biggest limitations of the study was the lack of motivation by the learners in both countries. Although learners received feedback on their writing competences, this was a scientific study with no immediate impact on grades or study prospects. As a result, a substantial percentage of responses could not be scored due either to lack of effort or to the kind of content they contained. Moreover, even though the kind of writing that is represented by the two TOEFL tasks is in line with expectations about writing proficiency as articulated in standards such as the Common European Framework of Reference for Languages, it is unclear how much time, exactly, teachers in the different classrooms spent on teaching this kind of writing, including how much time they spent on providing truly diagnostic feedback based on qualitative analyses of submitted responses outside of this study.

Similarly, while these tasks are didactically relevant to some degree, they are nevertheless isolated artificial assessment tasks in which responses are timed, collaboration with others is not possible, and the creation of multiple drafts is not advisable. Our study also underscores the need to try out writing prompts in terms of suitability for target populations, both in terms of content and linguistic difficulty. For example, difficulties in understanding the word "advertising" in the TV prompt seems to have led weaker students in the Swiss sample to stray off-topic sometimes. This, in turn, seemed to have impacted human rating distributions and led to differences in human – human agreement statistics between German and Swiss samples.

From an automated scoring perspective, construct representation within e-rater is certainly more sophisticated than it is in highly simplistic scoring algorithms, which would be more easily susceptible to the influence of essay length in

scoring, various types of cheating, and other mechanisms that result in fairness concerns for subgroups. However, an automated scoring engine of the type considered in this study will always remain somewhat limited relative to what human raters do in that its operationalization of the construct is componential rather than integrative. As a result, in operational TOEFL scoring, each response is always rated by at least one human rater, and human ratings for the integrated task currently receive twice the weight of machine scores; in addition, various advisory flags and filters are in place to detect validity threats.

Additional Implications

The findings in this research report demonstrate that a plug-and-play approach using an available automated scoring model always has potential risks associated with it in that acceptable performance on a new population is not guaranteed. At the same time, unless the context of assessment is truly high stakes, it is likely cost-prohibitive to go through various rounds of trialing prompts or tasks; collect human ratings; and then build, evaluate, and monitor automated scoring models. If use contexts are lower stakes for individual learners or larger administrative units, as is the case with many formative learning systems, then dynamic model building and evaluation is a smarter approach to scoring (e.g., Lochbaum, 2018). For example, multiple scoring models can be used to identify responses that require human ratings the most, such responses can be stored in an ever-growing database, and scoring models can be updated whenever predictive performance has increased notably.

Interestingly, no matter what kind of scoring architecture is developed, the start-up costs remain decidedly nontrivial. Automated scoring generally pays off only at large operational volumes and remains, at least at the moment, the purview of assessment and learning companies that intend to provide at-scale solutions. Moreover, the successful end-to-end design, implementation, and monitoring of such scoring systems requires the skill set and expertise of an interdisciplinary team comprising experts from areas such as natural language processing, computer science, language testing, psychometrics, data science, human scoring, assessment design, and learning sciences and involves myriad relevant design decisions that can lead to real validity threats (e.g., Rupp, 2018; see also Bejar, 2011; Bennett & Zhang, 2016; Williamson et al., 2012). Finding the appropriate balance of in-house investment and outsourced collaboration can be quite tricky.

Acknowledgments

The data analyzed in this research report came from the project "Measuring English Writing at Secondary Level" (MEWS), which was funded by the German Research Foundation (GZ: KO1513/12-1; AOBJ: 627083) and the Swiss National Science Foundation (grant 100019L_162675). We would like to thank our colleagues in natural language processing at ETS for their consultation around features and scoring approaches, notably Nitin Madnani, Swapna Somasundaran, and Jill Burstein, as well as Binod Gyawali for conducting all the model-building experiments. Thanks also go to our colleagues in the statistical analysis division for their help in processing data, organizing output, and preparing output files, most notably Chen Li and Duanli Yan. Furthermore, we want to thank the internal and external reviewers of this research report for improving its quality, with special thanks to Jakub Novak and Kim Fryer. On the international team, we would like to extend our thanks to Johanna Fleckenstein and Oliver Meyer, who were part of the research team and supported this effort through consultation and data analyses.

Notes

1 See http://www.ets.org/toefl for an overview of the test.

2 Human scores were included for responses that did not receive a score of 0 from any rater; other engine flags were not considered for case inclusion in these analyses.

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: John Wiley.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v. 2.0. *Journal of Technology, Learning, and Assessment, 4*(3), 1 – 30.

Autorengruppe Bildungsberichterstattung. (2018). *Bildung in Deutschland 2018. Ein indikatorengestützter Bericht mit einer Analyse zu Wirkungen und Erträgen von Bildung* [Education in Germany 2018: An indicator-supported report with an analysis regarding effects and outcomes of education]. Retrieved from https://www.bildungsbericht.de/de/bildungsberichte-seit-2006/bildungsbericht-2018

Beigman Klebanov, B., Madnani, N., Burstein, J., & Somasundaran, S. (2014). Content importance models for scoring writing from sources. In *Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (short papers)* (pp. 247–252). https://doi.org/10.3115/v1/P14-2041

Bejar, I. I. (2011). A validity-based approach to quality-control and assurance of automated scoring. *Assessment in Education: Principles, Policy, and Practice, 18*, 319–341.

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 142–173). New York, NY: Routledge.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay scoring: Current applications and future directions* (pp. 55–67). New York, NY: Routledge.

Dangeti, P. (2017). *Statistics for machine learning*. Birmingham, UK: Packt.

Deane, P. (2006). Strategies for evidence identification through linguistic assessment of textual responses. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 313–371). Mahwah, NJ: Erlbaum.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems, 9*, 155–161.

Fleckenstein, J., Keller, S., Krüger, M., Tannenbaum, R. J., & Köller, O. (2019). *Unlocking the TOEFL iBT rubrics for European writing assessment: Establishing a validity framework for cut scores in evidence-based standard setting*. Unpublished manuscript.

Haberman, S. J. (1996). *Advanced statistics: Description of populations*. https://doi.org/10.1007/978-1-4757-4417-0

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, NJ: Pearson/Prentice Hall.

Keller, S. (2016). Measuring English Writing at secondary level: Eine binationale Studie [Measuring English Writing at secondary level: A binational study]. *Babylonia, 3*, 46–48.

KMK. (2014). *National educational standards for English and French as foreign languages for the upper secondary track*. Cologne, Germany: Wolters.

Lochbaum, K. (2018, April). *Continuous flow: A hybrid of human and automated scoring*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Lomax, R. G., & Hahs-Vaughn, D. L. (2012). *Statistical concepts: A second course* (4th ed.). New York, NY: Routledge.

Lottridge, S. (2018, April). *Evaluating automated scoring feature and modeling upgrades relative to key criteria*. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, NY.

Madnani, N., Loukina, A., von Davier, A., Burstein, J., & Cahill, A. (2017). Building better open-source tools to support fairness in automated scoring. In *Proceedings of the first Workshop on Ethics in Natural Language Processing* (pp. 41–52). https://doi.org/10.18653/v1/W17-1605

Muthén, L. K., & Muthén, B. O. (2017). *Mplus* (Version 8.0) [Computer software]. Los Angeles, CA: Muthén and Muthén.

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of the e-rater scoring engine* (Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02158.x

Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education, 31*, 191–214. https://doi.org/10.1080/08957347.2018.1464448

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automated indexing. *Communications of the ACM, 18*, 613–620. https://doi.org/10.1145/361219.361220

Schweizerische Koordinationsstelle für Bildungsforschung. (2014). Bildungsbericht Schweiz *2014 [Educational report Switzerland 2014]*. Aarau, Switzerland: Author.

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. https://doi.org/10.4324/9780203122761

Somasundaran, S., Burstein, J., & Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In *Proceedings of COLING 2014: Technical papers* (pp. 950–961). Retrieved from http://aclweb.org/anthology/C14-1000

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wood, S. (2017, April). *Media, the public, and automated scoring*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Antonio, TX.

Yan, D., Rupp, A. A., & Foltz, P. (in press). *Handbook of automated scoring: Theory into practice*. New York, NY: Taylor and Francis Group/CRC Press.

Appendix A: Feature Description for Automated Scoring Models

Grammar, Usage, and Mechanics

These macrofeatures identify more than 30 error types, including errors in subject–verb agreement, preposition errors, pronoun errors, article errors, sentence fragments, missing commas, wrong word forms, and the like. These error types are summarized for each macrofeature as rescaled proportions of error rates relative to the essay length using the negative square root transformation.

Organization and Development

These macrofeatures are measures of text structure that automatically identify discourse categories for sentences in an essay: introductory material/background, thesis, main ideas, supporting ideas, and conclusion.

Collocations and Prepositions

This macrofeature is a measure of the correct use of collocations and prepositions and represents efforts to develop features capable of measuring positive indicators of writing in addition to those focusing on errors.

Average Word Length and Word Choice

These two macrofeatures measure lexical complexity using the average word length, as the name implies, and the median word frequency.

Sentence Variety

This macrofeature measures the diversity of the syntactic constructions of the sentences in an essay.

Content Vector Analysis Vocabulary (High Score) and Content Vector Analysis Vocabulary (Score Range)

There are two macrofeatures that measure the alignment of the content of an essay to the content of essays at the highest score point and each score point, respectively, using the methodology of *content vector analysis* (Salton, Wong, & Yang, 1975).

Discourse

This experimental macrofeature captures the organization of ideas in an essay. Based on the idea of lexical cohesion chains, it quantifies how ideas are initiated, continued, and terminated in essays. It also encodes how ideas are presented in relation to the discourse cues used to organize the text for the reader; see Somasundaran, Burstein, and Chodorow (2014) for more details.

Source-Use

This experimental macrofeature is specific to the TOEFL Integrated task and focuses on the use of information from the lecture relative to the reading passage. It comprises three microfeatures that quantify (a) how much of the material in the essay is drawn from the lecture stimulus, (b) how much of the material in the essay is drawn from the lecture stimulus as compared to from the reading passage text, and (c) how important the information from the lecture stimuli that the test taker used is. The macrofeature has two versions: a prompt-independent one, which we used for building generic models, and a prompt-specific one, which we used for building prompt-specific models; see Beigman Klebanov, Madnani, Burstein, and Somasundaran (2014) for more details.

Appendix B: Equations for Precision Statistics

Mean Square Error

$\mathrm{MSE}_H = \frac{1}{N}\sum\left(H_1 - H_2\right)^2$, the prediction error for human scores.

$\mathrm{MSE}_M = \frac{1}{N}\sum\left(H^* - M\right)^2$, the prediction error for machine scores

Proportional Reduction in Mean Square Error

$$\mathrm{PRMSE}_M = \frac{\left(C_{H,M}\right)^2}{V_T^* V_M}, \qquad \mathrm{PRMSE}_H = \frac{V_T}{V_H}$$

where

M = is bounded machine score

H_1 and H_2 = are observed human score from Rater 1 and Rater 2, respectively

$H^* = (H_1 + H_2)/2$, the average human score across Raters 1 and 2.

$V_R = \frac{\sum_{i=1}^{n}\left(H_{1i} - H_{2i}\right)^2}{2N}$, the variance of the human scores

$V_H = \mathrm{Var}(H^*)$, the variance of the average human scores

$V_M = \mathrm{Var}(M)$, the variance of the machine scores

$V_T = V_H - \frac{V_R}{2}$, the variance of the true scores

$C_{H,\,M} = \mathrm{Cov}(H^*, M)$, the covariance of average human scores and machine scores

Suggested citation:

Rupp, A. A., Casabianca, J. M., Krüger, M., Keller, S., & Köller, O. (2019). *Automated essay scoring at scale: A case study in Switzerland and Germany* (TOEFL Research Report No. RR-86). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12249

Action Editor: Anastassia Loukina

Reviewers: Aoife Cahill and Tim Davey