Applied Measurement in Education

ISSN: 0895-7347 (Print) 1532-4818 (Online) Journal homepage: http://www.tandfonline.com/loi/hame20

Establishing a crosswalk between the Common European Framework for Languages (CEFR) and writing domains scored by automated essay scoring

Mark D. Shermis

To cite this article: Mark D. Shermis (2018) Establishing a crosswalk between the Common European Framework for Languages (CEFR) and writing domains scored by automated essay scoring, Applied Measurement in Education, 31:3, 177-190, DOI: 10.1080/08957347.2018.1464451

To link to this article: https://doi.org/10.1080/08957347.2018.1464451

Accepted author version posted online: 16 Apr 2018.
Published online: 16 Apr 2018.

APPLIED MEASUREMENT IN EDUCATION
2018, VOL. 31, NO. 3, 177–190
https://doi.org/10.1080/08957347.2018.1464451

Routledge
Taylor & Francis Group

Establishing a crosswalk between the Common European Framework for Languages (CEFR) and writing domains scored by automated essay scoring

Mark D. Shermis

College of Education, University of Houston–Clear Lake, USA

ABSTRACT

This article employs the Common European Framework Reference for Language Acquisition (CEFR) as a basis for evaluating writing in the context of machine scoring. The CEFR was designed as a framework for evaluating proficiency levels of speaking for the 49 languages comprising the European Union. The intent was to impact language instruction so that "mastery" of one language has the same meaning as it does in another. A second objective is to provide a crosswalk for what one automated writing evaluation (AWE) system does in attending to the dimensions of the framework. The CEFR Framework is divided into five traits and different proficiency levels. The question then becomes: Does the AWE system attempt to measure these dimensions of writing? And, if so, how is this operationalized? Is it measuring aspects of communication that are not specified? The goal here is to create a common vocabulary between the writing community and those interested in AWE systems as to what is actually being measured by their software, and mapping that to a developmental scale of writing performance.

This article is a follow-up to Shermis' "future research" suggestion to use the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2002) as a possible basis for evaluating writing (Shermis, 2014). The CEFR, developed by the Council of Europe, was designed as a framework for evaluating proficiency levels of *speaking* for the 49 registered languages comprising the European Union. The intent was to impact language instruction so that "mastery" of one language has the same meaning as it does in another. Moreover, the philosophy underlying the framework embraces "plurilingualism" rather than "multilingualism" in that the goal is not to just master the mechanical aspects of speaking the candidate language, but also to comprehend the cultural aspects of understanding of what is being said or heard. This latter dimension is akin to observing a pianist who has learned the technical aspects of playing a piece, but lacks the passion and expression associated with a higher level of performance. The question here is: Can the CEFR be just as helpful in evaluating writing?

A second objective is to provide a crosswalk for what two automated writing evaluation (AWE) system do in attending to the dimensions of the framework. So, for example, if *fluency* and *cohesion* were identified as key elements in mastering a particular level of writing proficiency, the questions might be, "Does the AWE system attempt to measure these dimensions of writing? And, if so, how is this operationalized?" Is it measuring aspects of communication that are not specified? The goal here is to create a common vocabulary between the writing community and those interested in AWE systems as to what is actually being measured by their software, and mapping that to a developmental scale of writing performance. A lack of correspondence between the software and the

CONTACT Mark D. Shermis mshermis@uhcl.edu University of Houston–Clear Lake, 2700 Bay Area Blvd., Houston, TX 77058, USA

framework may suggest new areas for programming in software development or it may prompt the writing community to state more explicitly the desired trait being assessed.

The genesis of this work comes from the absence of a definition of what constitutes "good writing." For the purposes of this article, I use the following definition of writing:

> Writing is a medium of human communication that represents language and emotion through the inscription or recording of signs and symbols. In most languages, writing is a complement to speech or spoken language. Writing is not a language but a form of technology that developed as tools developed with human society. Within a language system, writing relies on many of the same structures as speech, such as vocabulary, grammar and semantics, with the added dependency of a system of signs or symbols. (Source: https://en.wikipedia.org/wiki/Writing)

The above definition suggests that there is great overlap between writing and speaking, but that there may be some differences. The following is a listing of some of the differences observed between speaking and writing:

- "Writing is usually permanent and written texts cannot usually be changed once they have been printed or written out.

Speech is usually transient, unless recorded, and speakers can correct themselves and change their utterances as they go along.

- A written text can communicate across time and space for as long as the particular language and writing system is still understood.

Speech is usually used for immediate interactions.

- Written language tends to be more complex and intricate than speech with longer sentences and many subordinate clauses. The punctuation and layout of written texts also have no spoken equivalent. However, some forms of written language, such as instant messages and e-mail, are closer to spoken language.

Spoken language tends to be full of repetitions, incomplete sentences, corrections, and interruptions, with the exception of formal speeches and other scripted forms of speech, such as news reports and scripts for plays and films.

- Writers receive no immediate feedback from their readers, except in computer-based communication. Therefore, they cannot rely on context to clarify things so there is more need to explain things clearly and unambiguously than in speech, except in written correspondence between people who know one another well.

Speech is usually a dynamic interaction between two or more people. Context and shared knowledge play a major role, so it is possible to leave much unsaid or indirectly implied.

- Writers can make use of punctuation, headings, layout, colors, and other graphical effects in their written texts. Such devices are not available in speech.

Speech can use timing, tone, volume, and timbre to add emotional context.

- Written material can be read repeatedly and closely analyzed, and notes can be made on the writing surface. Only recorded speech can be used in this way.
- Some grammatical constructions are only used in writing, as are some kinds of vocabulary, such as some complex chemical and legal terms.

Some types of vocabulary are used only or mainly in speech. These include slang expressions, and tags like *y'know, like*, etc." (http://www.omniglot.com/writing/writingvspeech.htm)

What is good writing?

The lack of a generally accepted definition of good writing is best exemplified by performing a Google Scholar search for the term and its alternatives (e.g., "good writing characteristics"). The result is a list of references that perhaps act as synonyms for the phrase such as "effective writing," "functional approach to the writing task," and a perennial favorite, "Good writing: I know it when I see it." The problem is that without a definition of good writing or an operational proxy, the phenomenon is impossible to accurately target and assess (Stiggins, 2007). Consequently, it becomes difficult to teach. Here is what the late Cleanth Brooks (2008) had to say in his popular *Fundamentals of Good Writing* text:

> There is no easy way to learn to write. There is no certain formula, no short cut, no bag of tricks. It is not a matter of memorizing rules or of acquiring a few skills. To write well is not easy for the simple reason that to write well you must think strait. And thinking strait is never easy.
> (p. 1)

Of course, the intent of the admonition was to make it clear that you will never achieve a good writing outcome if you do not have anything substantive to say. However, given this kind of advice, the average student might easily be discouraged, and think of writing as a hopeless enterprise.

In the search for an operational definition of good writing, most teachers of writing have turned to rubrics as a way to communicate what is expected of students. For the most part, the rubrics are based on characteristics of writing (traits) or components of writing (analytic scoring) (Stiggins & Bridgeford, 1983). For example, a popular rubric in the United States is the 6 + 1 Traits™ (Education Northwest, 1999). This rubric identifies six writing traits in addition to one presentation trait. These traits include: ideas, organization, voice, word choice, sentence fluency, *and* conventions.

The *Presentation* trait is sometimes used when the physical attractiveness of the writing might somehow affect its perception. When this is not the case, it is usually excluded as a dimension of scoring. The rubric is scored on a five-point scale that reflects the degree to which each of the dimensions is mastered. So, for instance, Table 1 lists the score point of 3 for Ideas.

The major drawback of this rubric, and others like it, is that while the parameters of each score point are rather clearly stated, the scores end up having relative meaning across grade levels. That is, a score of "3" for a third-grade writer is relatively different than a score of "3" from a ninth-grader. This does not render the rubric useless, but less functional than a scale that has an absolute definition across grade levels.

There have been attempts to create a developmental writing scale that would be invariant across grade levels, but with only limited success (Burdick et al., 2013). The *Writing Ability Developmental Scale*, developed by Metametrics uses an individual trait that is calibrated across multiple age levels from young writers to adults that is loosely based on their parallel work in Lexiles for reading

Table 1. A definition of "Ideas," a part of the 6 + 1 Traits™.

Ideas: The main message of the piece, the theme, with supporting details that enrich and develop that theme.

The paper has no clear sense of purpose or central theme. The reader must make inferences based on sketchy or missing details.

A. The writer is still in search of a topic
B. Information is limited or unclear or the length is not adequate for development
C. The idea is a simple restatement or a simple answer to the question
D. The writer has not begun to define the topic
E. Everything seems as important as everything else
F. The topic may be repetitious, disconnected, and contains too many random thoughts

Key Question: Did the writer stay focused and share original and fresh information or perspective about the topic?

Source. Education Northwest (1999)

(Burdick et al., 2013). It may be the case that while writing development is monotonic in nature, it may not be a linear scale (Shermis & Chang, 1997).

What is bad writing?

Most of the literature on bad writing focuses on problems that occur when one fails to communicate effectively. This includes features of writing that can lead to poor communication, including the use of passive voice, vagueness, wordiness, poor grammar and spelling, inappropriate language, stilted language, and lack of supporting information. For instance, many modern writers eschew the use of adverbs because they say it encourages "lazy writing." That is, it is easier to say "she went quickly" than to choose a more precise verb—*she hurried, she rushed, she hustled*. (http://upwritepress.com/BlogRetrieve.aspx?BlogID=2395&PostId=58094). Bad writing may contain grammatical, mechanical, usage, and stylistic errors, as well as errors of fact; may be inappropriate for an intended audience or purpose; or lack substance or entertainment value. The degree to which any or all of these things occur simultaneously contributes to how "bad" the writing might be assessed.

Taxonomies of writing

Most taxonomies of writing focus on the underlying purpose for the communication. For example, genre writing is based on rhetorical aspects of writing that underlie why the writer is generating the communication. Table 2 shows four commonly employed writing genres and their functions based on D'Angelo ("Modes of discourse," 1984).

Taxonomic labels for writing are not always mutually exclusive. So, within the frame of "Description," "Technical Writing" is a type of communication that attempts to clearly explain how things work. It is associated with operating manuals, business letters, and other artifacts where the goal is to convey information as plainly and efficiently as possible. Creative writing is thought to be outside the norms of professional, journalistic, academic, and technical writing. It may cut across all genres of writing and may fulfill multiple functions. Its typical function is to entertain.

The CEFR taxonomy

The CEFR was formally introduced in 2002 as a collective project of the European Union as a way to establish criteria and assessment procedures for evaluating individual language performance. The CEFR describes language proficiency in reading, writing, speaking, and listening on a six-level scale, clustered in three bands: A1-A2 (Basic User), B1-B2 (Independent User), and C1-C2 (Proficient User). It addresses three dimensions of communication: Understanding (Listening, Reading), Speaking (Interaction, Production), and Writing. The purpose of the CEFR is to create a system that has similar meaning across all languages so that a worker being classified as "Intermediate" in English has the same language skill set as another worker being classified as "Intermediate" in French or any of the other European Union–supported languages. The monotonic classifications include the

Table 2. Bain's forms of disclosure.

	Function	Subject	Organization	Language
Description	Evoke sense experience	Objects of senses	Space/time	Denotative and connotative, figurative, literal, impressionistic, objective
Narration	Tell a story, narrate an event	People and events	Space/time	As above
Exposition	Inform, instruct, present ideas	Ideas, generalizations	Logical analysis and classification	Denotative and factual
Argument	Convince, persuade, defend, refute	Issues	Deduction and induction	Factual and based on appeal

Note. Based on: Ruth and Murphy (1988) after D'Angelo (1976).

following categories: A1 (Breakthough or Beginner), A2 (Waystage or Elementary), B1 (Threshold or Intermediate), B2 (Vantage or Upper Intermediate), C1 (Effective Operational Proficiency or Advanced), and C2 (Mastery or Proficiency). Table 3 lists each level and some "can-do" statements of individuals who are classified at that level.

In the CEFR Framework, writing is a core component of the assessment along with speaking and understanding. Writing proficiency is divided into five traits (range, accuracy, fluency, interaction, coherence) and six different proficiency levels. The five traits are not explicitly defined, but their levels are described in the CEFR self-assessment grid listed in Table 4.

Range refers to the variety of topics that one can write on and audiences that one can write for. From an instructional standpoint, increasing the range of writing is designed to enhance writing *fluency*.

Accuracy refers to how correctly the writer employed language, including their use of grammar and vocabulary. For writing purposes where the correctness of the information is important, accuracy refers to how closely the response approximates a modeled answer.

Fluency is the ability of an individual to deliver information quickly and with expertise. Peter Elbow, in his classic text *Writing without Teachers*, describes the importance of fluency in the instruction and assessment of writing (Elbow, 1973). His developmental model is geared to encouraging the writer to produce more text and then edit it. Taking a cognitive perspective on writing, McCutchen, Teske, & Bankston (2008) write, "Fluent text production can influence the writing process both directly and indirectly because inefficient text production can consume [cognitive] resources that might otherwise be devoted to higher level processes such as planning and revising" (p. 457).

Interaction refers to the amount and level of communication between the writer and his or her audience. This might be a long passage, hypothetical question, or a sharing of others' ideas.

Coherence refers to the characteristic of a communication that makes it semantically meaningful. The linguistic elements that make a text coherent are subsumed under the term "cohesion." A coherent text provides information and context in a manner that is both efficient and familiar to the audience.

So at the A1 (most basic) level of *coherence*, the writer is supposed to be able to "link words or groups of words with very basic linear connectors like 'and' or 'then.'" At the highest level of *coherence* functioning, a writer can summarize and critique professional or literary work.

Good writing redux

So what can be made of all of this? Here is one way to define good writing within the context of the CEFR:

> Writing is a medium of human communication that represents language and emotion through the inscription or recording of signs and symbols. Good writing is evaluated through the developmental progression of five dimensions of communication, including range, accuracy, fluency, interaction, and coherence. The weights given to each of the five dimensions are influenced by the purpose of the written communication, sometimes referred to as writing genre for which there are a number of taxonomies. So for instance, *technical writing* may place a premium on coherence and range while *creative writing* may emphasize fluency and interaction. For any genre of writing there are minimum dimensional thresholds of acceptability, but it is the dimensional emphasis that shifts from genre to genre.

The challenge for assessment of writing is to determine the appropriate weights for each genre, ascertain whether these change from developmental step to developmental step, and formulate minimum acceptable thresholds. For machine scoring, an additional challenge is to ensure that there is an appropriate crosswalk between the CEFR dimensions and the proxies that are used by the machine scoring algorithms to make score predictions.

Table 3. Common reference levels: Self-assessment grid. Source: Council of Europe, 2002.

		A1	A2	B1
UNDERSTANDING	Listening	I can recognize familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly.	I can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (e.g., very basic personal and family information, shopping, local area, employment). I can catch the main point in short, clear, simple messages and announcements.	I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, and so on. I can understand the main point of many radio or TV programs on current affairs or topics of personal or professional interest when the delivery is relatively slow and clear.
	Reading	I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues.	I can read very short, simple texts. I can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short simple personal letters.	I can understand texts that consist mainly of high frequency every day or job-related language. I can understand the description of events, feelings and wishes in personal letters.
SPEAKING	Spoken Interaction	I can interact in a simple way provided the other person is prepared to repeat or rephrase things at a slower rate of speech and help me formulate what I'm trying to say. I can ask and answer simple questions in areas of immediate need or on very familiar topics.	I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. I can handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself.	I can deal with most situations likely to arise whilst travelling in an area where the language is spoken. I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g., family, hobbies, work, travel, and current events).
	Spoken Production	I can use simple phrases and sentences to describe where I live and people I know.	I can use a series of phrases and sentences to describe in simple terms my family and other people, living conditions, my educational background and my present or most recent job.	I can connect phrases in a simple way in order to describe experiences and events, my dreams, hopes and ambitions. I can briefly give reasons and explanations for opinions and plans. I can narrate a story or relate the plot of a book or film and describe my reactions.
WRITING	Writing	I can write a short, simple postcard, for example sending holiday greetings. I can fill in forms with personal details, for example entering my name, nationality and address on a hotel registration form.	I can write short, simple notes and messages relating to matters in areas of immediate need. I can write a very simple personal letter, for example thanking someone for something.	I can write simple connected text on topics which are familiar or of personal interest. I can write personal letters describing experiences and impressions.
		B2	**C1**	**C2**
UNDERSTANDING	Listening	I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programs. I can understand the majority of films in standard dialect.	I can understand extended speech even when it is not clearly structured and when relationships are only implied and not signaled explicitly. I can understand television programs and films without too much effort.	I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided I have some time to get familiar with the accent.
	Reading	I can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. I can understand contemporary literary prose.	I can understand long and complex factual and literary texts, appreciating distinctions of style. I can understand specialized articles and longer technical instructions, even when they do not relate to my field.	I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialized articles and literary works.
SPEAKING	Spoken Interaction	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my views.	I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. I can handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself.	I can deal with most situations likely to arise whilst travelling in an area where the language is spoken. I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel, and current events).
	Spoken Production	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.	I can present clear, detailed descriptions of complex subjects integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.	I can present a clear, smoothly flowing description or argument in a style appropriate to the context and with an effective logical structure which helps the recipient to notice and remember significant points.
WRITING	Writing	I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences.	I can express myself in clear, well-structured text, expressing points of view at some length. I can write about complex subjects in a letter, an essay or a report, underlining what I consider to be the salient issues. I can select style appropriate to the reader in mind.	I can write clear, smoothly flowing text in an appropriate style. I can write complex letters, reports or articles which present a case with an effective logical structure which helps the recipient to notice and remember significant points. I can write summaries and reviews of professional or literary works.

Table 4. Common reference levels: Qualitative aspects of spoken language use. Source: Council of Europe, 2002.

	Range	Accuracy	Fluency	Interaction	Coherence
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g., in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using nonverbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turn-taking, referencing, allusion making, and so on.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organizational patterns and a wide range of connectors and other cohesive devices.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skillfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organizational patterns, connectors and cohesive devices.
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause mis-understanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions. There are few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
A2	Uses basic sentence patterns with memorized phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connectors like "and," "but," and "because."
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorized repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like "and" or "then."

Criticisms of the CEFR

While the CEFR has enjoyed relatively good acceptability in Europe as a communications framework and has been applied to assessment in the United States (Tannenbaum & Wylie, 2008), it has certainly not escaped criticism. Some of the criticism is specific to the CEFR and some could be applied to any performance scale. Weir (2005) probably has the best summary of these:

(1) the scales are premised on an incomplete and unevenly applied range of contextual variables/performance conditions (context validity);
(2) little account is taken of the nature of cognitive processing at different levels of ability (theory-based validity);
(3) activities are seldom related to the quality of actual performance expected to complete them (scoring validity);
(4) the wording for some of the descriptors is not consistent or transparent enough in places for the development of tests.

The first criticism is basically a recognition that the CEFR cannot be all things to all people. The genesis of the CEFR was to develop language performance classifications that would allow employers to reliably ascertain the language skills sets of prospective employees. The scales assume that development of language skill sets is roughly equivalent across the 49 languages that are covered in the European Union. In this sense, the variables associated with each classification may differ, the categories themselves may only have ordinal properties, and the framework may not lend itself as well to a specific testing purpose or form of testing (e.g., diagnostic testing).

The second criticism suggests that the CEFR is not necessarily aligned with writing development theory or that the assessments based on the CEFR would be consistent with the course of writing development. Most instructors of writing will encourage students to plan their writing, generate a draft of the document, obtain feedback, and revise, and perhaps iterate these last three steps several times. This developmental approach usually takes some considerable time that may not be available for CEFR assessments. Moreover, the methodology used for CEFR performance tasks may be incongruent with performance per se. That is, relying on a multiple-choice test to assess writing may be fundamentally at odds with trying to demonstrate how one actually writes.

The third criticism has to do with the scoring context of the performance assignment. The question here is, "To what degree do the tasks on a CEFR assessment mimic the context of a real-life situation?" Is describing a patient's condition in an emergency room similar to that in a more artificial setting? Is the pressure the same? The distractions? The environment? Basically, this is the difference, even with a performance assessment, between the constraints of a test environment compared to that of a real-life setting. The acknowledgment is that actual performance is likely to deviate from that on a test item seeking to assess that performance.

The last criticism has to do with the lack of precision taken on by the classification labels of the CEFR categories. There is no collectively agreed-on cutoffs for categories such as "Threshold" and "Waystage" or even agreement that these are good descriptors. In order to address this challenge, the CEFR has a list of "can do" statements that operationally define the categories. Table 5 shows a list of "can do" statements for the CEFR. However, even with these "can do" statements, there may be specific tasks aligned within a category that are harder or easier than tasks in adjacent categories. This issue arises in other curricular areas as well. For example, there are a few tasks in calculus that are functionally easier than some tasks in trigonometry even though the former domain is considered harder than the latter (Shermis & Chang, 1997).

Table 5. Performance examples of CEFR writing. Source: Council of Europe, 2002.

Level	Writing
A1	I can write a short, simple postcard, for example sending holiday greetings. I can fill in forms with personal details, for example entering my name, nationality and address on a hotel registration form.
A2	I can write short, simple notes and messages relating to matters in areas of immediate need. I can write a very simple personal letter, for example thanking someone for something.
B1	I can write simple connected text on topics that are familiar or of personal interest. I can write personal letters describing experiences and impressions.
B2	I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences.
C1	I can express myself in clear, well-structured text, expressing points of view at some length. I can write about complex subjects in a letter, an essay or a report, underlining what I consider to be the salient issues. I can select style appropriate to the reader in mind.
C2	I can write clear, smoothly flowing text in an appropriate style. I can write complex letters, reports, or articles that present a case with an effective logical structure that helps the recipient to notice and remember significant points. I can write summaries and reviews of professional or literary works.

Formulating a crosswalk

To this point an argument has been put forth for a definition of good writing and employing the CEFR as a possible way to operationalize that definition. The next step would be to formulate a crosswalk to determine how current AES systems align with the CEFR in any functional way, and then try to resolve whether the machine engine proxies reasonably address the domain defined by the CEFR trait. For instance, if the machine engine is calculating a measure of *lexical complexity*, does this match well with the definition of the CEFR trait *fluency* or is it a component of some other trait? Is the one measure sufficient coverage or might there be elements of the trait that are absent or under-represented? Using the example above, Elbow (1998) suggests that the length of the communication influences how fluent the communication can be. If one is not saying much, the communication cannot be fluent. Does *lexical complexity* incorporate elements of length in its definition? These are the issues that need to be teased out in formulating a crosswalk. Once completed, it could be that there are missing areas or there may be an overabundance of proxies in one CEFR trait.

An example

In order to demonstrate the crosswalk, alignments from two automated essay scoring engines are illustrated—e-rater® and Constructed-Response Automated Scoring Engine(CRASE™). E-rater is an automated essay evaluation and scoring system that was developed by researchers at the Educational Testing Service (ETS; Attali & Burstein, 2006; Burstein, Tetreault, & Madnani, 2013). E-rater first became operational in 1999 when it was deployed to provide one of two scores for essays on the writing section of the Graduate Management Admissions Test, a high-stakes, large-scale assessment. In conjunction with human ratings, e-rater is also presently used for the computer-based Test of English as a Foreign Language® iBT^{TM} and as a check score for the Graduate Record Examination. In the *Criterion*® Online Writing Evaluation service—a Web-based writing tool that helps students plan, write, and revise their essays—e-rater is used to foster best instructional and assessment practices (Burstein, 2012).

One of a family of scoring engines used at ETS, e-rater aligns a defined writing construct with natural language processing methods in order to identify linguistic features in student and test-taker writing for the purpose of scoring and evaluation (e.g., diagnostic feedback). Consideration of aspects of the writing construct in earlier system development (Attali & Burstein, 2006; Burstein, Tetreault, & Madnani, 2013) and more recent research and development support the enhancement of the system's construct coverage with regard to the structure of argumentation (Beigman-Klebanov, Madnani, & Burstein, 2013), discourse coherence (Burstein, Tetreault, Chodorow, Blanchard, & Andreyev, 2013b), and vocabulary usage (Beigman-Klebanov et al., 2013). Using statistical and rule-based natural language processing methods, the e-rater software currently identifies and extracts several feature classes for model building and essay scoring (Attali & Burstein, 2006; Burstein, Chodorow, & Leacock, 2004; Burstein, Tetreault, & Madnani, 2013).

Feature development and re-evaluation is dynamic, and specific feature variables may vary as the system is updated with new releases. Individual feature classes represent an aggregate of multiple features. The feature classes—the variables of writing that constitute the construct model developed for score prediction—include the following: (a) grammatical errors (e.g., *subject–verb agreement errors*), (b) word usage errors (e.g., *their* versus *there*), (c) errors in writing mechanics (e.g., *spelling*), (d) presence of essay-based discourse elements (e.g., *thesis statement, main points, supporting details*, and *conclusions*), (e) development of essay-based discourse elements, (f) style weaknesses (e.g., *overly repetitious use of vocabulary*), (g) two content vector analysis-based features to evaluate topical word usage, and (h) a feature that considers *correct usage* of prepositions and collocations (e.g., *powerful computer* vs. *strong computer*) (Futagi, Deane, Chodorow, & Tetreault, 2008), and sentence variety. The set of features in (h) represent positive features, rather than errors in conventions. Because proper usage of English prepositions and collocations is especially difficult for English learners, the addition of these features also expands e-rater's ability to recognize characteristics of writing important for assessing nonnative writers. Figure 1 illustrates the breakdown of e-rater. More details about specific features aggregated within a feature class may be found in Attali and Burstein (2006).

Human-assigned holistic scores are used to build *e-rater* models. A randomly selected training sample of at least 250 human-scored essays is processed through *e-rater*, which extracts the features described above. Features are aggregated into conceptually related groups and converted to a vector (list) of numerical feature values. Using a regression modeling approach, the values from this sample are used to determine an appropriate weight for each feature (Attali & Burstein, 2006; Burstein, Tetreault, & Madnani, 2013; Davey, 2009). To score a new, unseen essay during a test administration, the same process is performed vis-à-vis feature extraction, and conversion of features to a vector format. To compute the final score prediction, these values are then multiplied by the weights associated with each feature, and a sum of the weighted feature values is computed (Attali,

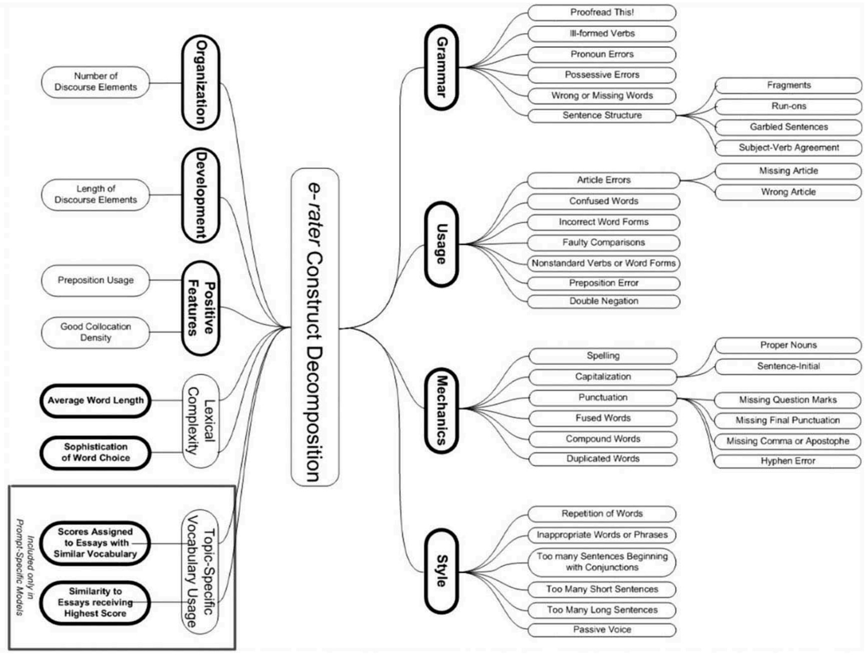

Figure 1. Organization and construct coverage of e-rater v10.1.

Bridgeman, & Trapani, 2010). In addition to providing holistic scores using construct-aligned language features, e-rater provides advisories for detecting such construct-irrelevant elements and attempts to enhance the overall score (Higgins, Burstein, & Attali, 2006). Many features used for e-rater scoring are also used for feedback in *Criterion*. These include the grammar, usage, mechanics, and styles features, as well as the organization and development features described above.

Pacific Metrics' automated scoring engine, CRASE™, scores responses to items typically appearing in large-scale assessments: (a) essay length writing prompts; (b) short answer constructed response items in mathematics, English language arts, and science; (c) math items eliciting formulae or numeric answers; and, (d) technology-enhanced items (e.g., Drag and Drop, Graphing). It has been used in both formative and high-stakes summative assessments, providing rapid turnaround and delivering cost savings over traditional hand scoring methods. The system is highly customizable, both in terms of the configurations used to build machine scoring models and in terms of the how the system can blend human scoring and machine scoring (i.e., hybrid models). CRASE is a fully integrated Java-based application that runs as a Web service. The term integrated refers to its ability to: (a) score any of several different item types as a single software application, (b) interface with Web-based assessment delivery platforms for immediate turnaround of scores, and (c) integrate with vendor-based electronic hand scoring systems for monitoring or dual scoring.

For the scoring of writing prompts such as those appearing in this study, the feature extraction step is organized around the 6 + 1 Trait® Model, a product of Education Northwest (http://educationnorthwest.org/traits) that is used in some form by many states for K–12 writing applications. The 6 + 1 Trait Model conceptualizes six traits of writing (ideas, sentence fluency, organization, voice, word choice, and conventions) along with the "+1," which is "written presentation." Written presentation, as outlined in the 6 + 1 Trait model, is not assessed by CRASE. To extract features related to the Ideas trait, CRASE uses a set of bag-of-words methodologies such as naïve Bayes classifiers and word vector analyses. For the Sentence Fluency trait, basic sentence characteristics are collected (e.g., average sentence length) as well as an entropy measure of sentence variety that takes into consideration sentence complexity and sentence type (e.g., exclamatory sentences such as "Wow!"). For Organization, CRASE extracts paragraphing statistics, thesis/conclusion identifiers, as well as counts of various discourse phrasing markers. To extract features related to Voice, CRASE extracts the use of terms representing informal language, tone, level of personal engagement with the topic, and the use of over-used words (e.g., very). For the Word Choice trait, the CRASE system extracts features related to word uniqueness, word complexity as well as an overall measure representing part of speech usage in the essay compared to high-scoring essays in the training set. Finally, to extract features related to Conventions, CRASE collects data on spelling errors as well as usage and mechanics errors.

CRASE is actually comprised of a number of scoring modules, including the essay scoring engine described above. The engine can also combine multiple modules for scoring and include scores from one module into another. For example, the system can call its content scoring engine to look for the presence or absence of concepts in essays and then submit those values into the essay scoring module. This process was used for some items in this study.

Once the features are collected from the various scoring modules, then CRASE uses the feature values to generate a scoring model to predict scores using linear or logistic regression based on the scored training sample. Bayesian priors can also be employed.

Table 6 shows the crosswalk between the CEFR traits and the feature sets employed by e-rater and CRASE. It should be noted that as of this writing only one automated essay scoring engine—Write

Table 6. CEFR crosswalk for e-rater® and CRASE™.

CEFR Trait	Range	Accuracy	Fluency	Interaction	Coherence
e-rater Construct	Mechanics Style	Grammar Topic Specific Vocabulary Usage	Lexical Complexity	Usage Positive Features	Organization Development
CRASE Construct	Usage and Mechanics	Word Choice	Sentence Fluency	Voice	Discourse Phrasing

and Improve™—explicitly uses the CEFR framework for interpreting score output. However, the developers of this engine have not yet published a proxy list of variables or constructs that it uses to define score predictors. At first glance there are some commonalities between e-rater and CRASE with regard to their match-up with the CEFR traits. For example, the CEFR trait *range* lines up with both e-rater and CRASE on the *mechanics* proxy. However, CRASE bundles *usage and mechanics* whereas e-rater splits these proxies out into separate categories. *Usage* with e-rater is more aligned with the CEFR trait of *interaction*. The CEFR Trait of *accuracy* is covered by e-rater with the domain areas of *grammar* and the construct *topic specific vocabulary usage*. This latter term applies to vocabulary that are related to the content of the material covered by the essay, if it is content related. For CRASE this trait is addressed by the construct of *word choice*. E-rater uses *lexical complexity* to match the CEFR trait of *fluency* and CRASE uses the construct *sentence fluency*. For the CEFR trait of interaction, e-rater employs both *usage* and *positive features* while CRASE opts for the 6 + 1 Traits-inspired construct of *voice*. Finally, the CEFR trait of *coherence* is mapped by two e-rater features—*organization* and *development*, but only one CRASE feature—*discourse phrasing*.

Discussion

At first glance, it would seem that the naming conventions between the CEFR traits and their machine scoring counterparts are only loosely tied together, but upon further consideration there are some commonalities. For instance, when it comes to operationalizing the trait *range*, the machine systems refer to this as *mechanics* and *style. Accuracy* is essentially defined as *word choice; fluency* as *sentence fluency; interaction* as *usage*; and *coherence* as *organization and development*. The clearest alignments seem to be with the CEFR trait of *fluency*, followed by *coherence*, and perhaps *accuracy*, although this conclusion may be overly simple. For example, the functional equivalent for *fluency* in e-rater is *lexical complexity* which is probably just a subset of *fluency*. Part of the challenge may be that machine measures of *fluency* depend on length counts, and some developers of machine scoring software want to ensure that their predictions avoid having the appearance of relying on so-called "superficial" variables. They may deliberately exclude certain variables (e.g., sentence length) to avoid this problem. The CEFR traits of *range* and *interaction* seem to be less well aligned either because they are conceptually more ambiguous or because they are harder to operationalize.

So where do we go from here? Both the fields of writing and machine scoring could benefit from adopting CEFR traits as a way to converge on a common vocabulary. For writing, the next step would be to ascertain a series of agreed-on empirical markers corresponding to CEFR thresholds. Aside from the production of specific artifacts, are there specific markers between the thresholds (e.g., beginner to waystage) that would provide a mechanism to create reliable cut scores? Several tests purportedly do just this, but each is unique in the way the cut scores are implemented. What is proposed here is a set of comparability studies for assessments that use the CEFR as an outcome framework.

With regard to the machine scoring community, it would be helpful to better align and flesh out their operationalizations of the CEFR trait domains. From the example above, e-rater defines *lexical complexity* as a weighted combination of *average word length* and *sophistication of word choice*. It is unlikely that these two proxies would fully define what is meant by fluency (nor were they designed to), so the task would be to describe how proxies used in e-rater or any machine scoring system would align with the CEFR traits, and to identify gaps in coverage. These gaps may be deliberate or simply reflect the state-of-the-art with respect to program capabilities for machine scoring. The gaps then become targets for future development.

The purpose of this article was to encourage both the writing and machine scoring communities to begin to work together so that writing assessments can be more accurate, more frequent, of greater utility, and reflect agreed-on traits. Part of the challenge is devising a definition of "good writing" and formulating a framework within which both writers and those who evaluate writing can operate. In this article, I have proposed an operational definition of "good writing" and advocated for the CEFR traits as a possible framework for conveying it to writers. By using a common vocabulary and working to more comprehensively operationalize the five CEFR traits, both communities can begin having constructive conversations about what machine scoring is and isn't addressing in the various programs that have been developed. Until the communities can agree on what they are measuring, they will have an impossible task of evaluating how well it is being measured.

ORCID

Mark D. Shermis http://orcid.org/0000-0002-0463-572X

References

Attali, Y., Bridgeman, B., & Trapani, C. S. (2010). Performance of a generic approach in automated essay scoring. *Journal of Technology, Learning, and Assessment*, *10*. Retrieved from http:jtla.bc.edu

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, *4*. Retrieved from http:jtla.bc.edu

Beigman-Klebanov, B., Madnani, N., & Burstein, J. (2013, January 1). *Using pivot-based paraphrasing and sentiment profiles to improve a subjectivity lexicon for essay data*. Retrieved July 17, 2017, from http://www.transacl.org/papers/

Brooks, C. (2008). *Fundamentals of good writing*. New York, NY: Fitts Press.

Burdick, H., Swartz, C. W., Stenner, A. J., Fitzgerald, J., Burdick, D., & Hanlon, S. T. (2013). Measuring students' writing ability on a computer-analytic developmental scale: An exploratory validity study. *Literacy Research and Instruction*, *52*(4), 255–280. doi:10.1080/19388071.2013.812165

Burstein, J. (2012). Fostering best practices in writing instruction and assessment with E-rater®. In N. Elliot & L. Perelman (Eds.), *Writing assessment in the 21st century essays in Honor of Edward M. White*. New York: Hampton Press. doi:10.1002/j.2333-8504.2011.tb02250.x/full

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

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater automated scoring system. In M. D. Shermis & J. Burstein (Eds.), Handbook of Automated Essay Evaluation (pp. 55–67). New York, NY: Routlege.

Burstein, J., Tetreault, J., Chodorow, M., Blanchard, D., & Andreyev, S. (2013). *Automated evaluation of discourse coherence quality in essay writing*. In M. D. Shermis & J. Burstein (Eds.), Handbook of Automated Essay Evaluation (pp. 267–280). New York, NY: Routledge.

Council of Europe. (2002). *Common European framework of reference for languages*. Strausburg Cedex, France: Author.

Davey, T. (2009). Principles for building and evaluating e-rater models. Presented at the National Council on Measurement in Education, San Diego, CA.

D'Angelo, F. J. (1984). Nineteenth-century forms/modes of discourse: *A critical inquiry. College Composition and Communication*, 35(1), 31–42.

Education Northwest. (1999). *6+1 traits of writing rubric*. Retrieved December 1999, from http://educationnorthwest.org/traits

Elbow, P. (1973). *Writing without Teachers*. New York, NY: Oxford University Press.

Elbow, P. (1998). *Writing with power: Techniques for mastering the writing process*. New York, NY: Oxford University Press.

Futagi, Y., Deane, P., Chodorow, M., & Tetreault, J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, *21*, 353–367. doi:10.1080/09588220802343561

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering*, *12*(2), 145–159. doi:10.1017/S1351324906004189

McCutchen, D., Teske, P., & Bankston, C. (2008). Writing and cognition: *Implications of the cognitive architecture for learning to write and writing to learn*. In C. Bazerman (Ed.), Handbook of Research on Writing History, Society, School, Individual, Text (pp. 451–470). New York, NY: Lawrence Erlbaum Associates

Ruth, J., & Murphy, S. M. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: A United States demonstration and competition, results, and future directions. *Assessing Writing*, *20*, 53–76. doi:10.1016/j.asw.2013.04.001

Shermis, M. D., & Chang, S.-H. (1997). The use of item response theory (IRT) to investigate the hierarchical nature of a college mathematics curriculum. *Educational and Psychological Measurement*, *57*(3), 450–458. doi:10.1177/0013164497057003006

Stiggins, R. (2007). *An introduction to student-involved assessment for learning* (5th ed.). Portland, OR: Assessment Training Institute.

Stiggins, R. J., & Bridgeford, N. J. (1983). An analysis of published tests of writing proficiency. *Educational and Psychological Measurement*, *2*, 6. doi:10.1111/emip.1983.2.issue-1

Tannenbaum, R. J., & Wylie, E. C. (2008). Linking English-language test scores onto The Common European Framework of Reference: An application of standard-setting methodology. *ETS Research Report Series*, *2008*, i–75. doi:10.1002/j.2333-8504.2008.tb02120.x

Weir, C. J. (2005). Limitations of The Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*(3), 281–300. doi:10.1191/0265532205lt309oa