

***Automated diagnostic writing tests: Why? How?***

Elena Cotos and Nick Pendar

Technology for Second Language Learning Conference

2007

Iowa State University

## **1. Introduction**

The importance of diagnostic assessment is undeniable since diagnostic tests are “closest to being central to learning” a second or foreign language (Alderson, 2005, p. 4). However, although diagnosis of language proficiency is an old topic, it still lacks a clear theoretical understanding, is under-investigated, and therefore underrepresented in the field of language testing. Despite the myriad reasons that hinder progress in diagnostic assessment, this area can gain a lot of potential if it embraces current technological advancements. In this paper, we will argue that Computer-Assisted Language Testing (CALT) would benefit from employing automated scoring systems, used in high-stakes standardized tests, for the purpose of diagnosing learner written language proficiency. To justify this point, we will first reflect on what a computer-assisted diagnostic writing test should be, what is to be diagnosed in learner writing, and why constructed responses, as opposed to multiple choice, for instance, are more suitable for language diagnosis. Since Automated Essay Scoring (AES) systems evaluate constructed responses, we will then closely examine AES programs and techniques to show how they could potentially yield diagnostic information about the constructs of interest and provide complex individualized feedback. We will conclude with deliberations on the potential advantages of AES systems for diagnostic language assessment and, in the long run, for language learning.

## **2. Computer-based diagnostic tests: what they should be**

According to the age-old definition, diagnostic language tests aim at identifying learners’ areas of “strengths and weaknesses” (Alderson et al., 1995; Bachman and Palmer, 1996; Davies et al., 1999; Moussavi, 2002) in order to help improve learning. The strengths identified should point to the level a learner has reached, and the weaknesses detected should indicate areas for improvement. On a theoretical level, for such diagnosis to happen, evidence on the development of second/foreign language proficiency is needed. As Alderson (2005) puts it, “[w]ithout a theory of development, a theory, perhaps also, of failure, and an adequate understanding of what underlies normal development as well as what causes abnormal development or lack of development, adequate diagnosis is unlikely” (p. 25). Undoubtedly, language development theory is extremely important in language testing for purposes of construct definition and level scale generation; however, it is still a major concern in the field of Second Language Acquisition (SLA) and will not be addressed in this paper. What we are concerned with at this point is the practical level of diagnosing learner language proficiency. From this perspective, for a diagnostic test to be truly valuable, detailed analysis and report on the learner’s performance is imperative. Since obtaining detailed profiles of learner written performance across various components of the construct for the purpose of diagnosis writing proficiency appears to be possible in CALT, it will be considered in this discussion.

In order to develop a computer-based diagnostic writing test, one must be aware of its inherent characteristics. Existing CALT literature provides very limited direction as to how diagnostic tests must be constructed, conducted, and validated; consequently, valid diagnostic language tests are extremely scarce. In most cases they are not genuine since they are used for placement purposes or are replaced by proficiency or achievement

tests simply because “virtually any test has some potential of providing diagnostic information” (Bachman, 1990, p. 60). What distinguishes a genuine diagnostic test from other types of assessment is self-referencing. In achievement and norm-referenced tests, for instance, referencing is typically with respect to a population, while “in a diagnostic test the student’s performance is compared against his or her expected performance” (Bejar, 1984, p. 176). Furthermore, a diagnostic test should be oriented towards learning by providing students with explicit feedback to be acted upon in addition to displaying immediate results. It should generate a detailed analysis of learner responses, which should lead to remediation in instruction. Irrespective of the kind of instruction, it can be based on the content that has been or will be covered in the teaching process and become an essential part of individualized instruction or self-instruction. Unlike many other tests, its results should be qualitative or analytic rather than quantitative, and their interpretation should not be used for high-stakes decisions. Finally, it is advised that a diagnostic test should not impose time limits (Schonell and Schonell, 1960) because the test-taking conditions are not part of the construct, as it will be described further.

### **3. Constructs in computer-based diagnostic writing tests**

What should a diagnostic test evaluate to ensure that the learner’s strengths and weaknesses are revealed? This question is very difficult to answer, but it is unavoidable because defining the construct is essential in testing. It requires a reliable framework, which would describe proficiency development at different levels. Although not entirely supported by SLA research, a number of such frameworks have been elaborated; e.g., ACTFL scales (American Council for Teaching of Foreign Languages, 1983), International Language Proficiency scales (Wylie and Ingram, 1995/1999), Canadian Benchmarks (Pawlikowska-Smith, 2000), and the Common European Framework of Reference CEFR (Council of Europe, 2001). A first attempt to follow one of these scales is DIALANG – a unique piloting effort to develop and implement computer-based diagnostic tests. This European Union-funded project attempts to provide diagnostic information about learners’ reading, listening, writing, grammar, and vocabulary proficiency in 14 languages relying on CEFR. Since it is the only diagnostic test that is exploring empirical results, we will consider what it is that should be diagnosed in the light of how DIALANG defines its constructs. In particular, due to the specific focus of this paper, we will cover the construct components that can be diagnosed from learners’ writing.

The main aspects that are targeted by DIALANG are *textual organization*, *appropriacy*, and *accuracy* in writing for communicative purposes such as providing information, arguing a point, or social interaction. For textual organization, learners are diagnosed based on how good they are at detecting coherence and cohesion markers; for appropriacy, based on how well they can set the tone and the level of formality in the text; and for accuracy, based on how they can cope with grammar and mechanics. For the latter, Alderson (2005) provides a somewhat detailed<sup>1</sup> frame of grammatical structures.<sup>2</sup>

---

<sup>1</sup> The frame is “somewhat detailed” considering that it was meant to inform item development for 14 languages DIALANG covers. More details were added depending on the peculiarities of individual languages.

<sup>2</sup> Alderson (2005) discusses grammatical categories when describing DIALANG’s grammar test; however, we found this material very relevant in this context.

Table 1. Morphological and Syntactical categories

<i>Morphology</i>	<i>Syntax</i>
<i>Nouns</i> : inflection – cases definite/indefinite – articles proper/common <i>Adjectives and Adverbs</i> : inflection comparison	<i>Organization/Realization of Parts of Speech</i> : word order – statements, questions, exclamation agreement <i>Simple and Complex Clauses</i> : coordination subordination deixis
<i>Pronouns</i> : inflection context <i>Verbs</i> : inflection – person, tense, mood active/passive voice <i>Numerals</i> : inflection context	<i>Punctuation</i>

Assessment of writing proficiency would be incomplete without an analysis of learner’s vocabulary. DIALANG incorporates separate vocabulary tests, which are targeted at learners’ knowledge of the meanings of single words and word combinations. Specifically, knowledge of vocabulary is evaluated from several perspectives – word formation by affixation and compounding; semantic ties between synonyms, antonyms, hyponyms, polysemantic words, etc.; word meanings including denotation, connotation, semantic fields; and word combinations such as idioms and collocations.

#### 4. Computer-based diagnostic writing test items

To elicit information about the construct of interest, the area of CALT has made use of test items of various levels of constraint. Scalise and Bernard (2006) provide a comprehensive taxonomy for electronic assessment questions and tasks that include multiple choice, selection/identification, reordering/rearranging, substitution/correction, completion, construction, and presentation/portfolio. Existing diagnostic tests, however, still follow the limited indirect testing approach, in which components of the construct are assessed “indirectly through traditional objectively assessable techniques like multiple choice” (Alderson, 2005, p. 155). Indeed, our example, DIALANG, consists of such item formats as multiple choice, drop-down menus, text-entry, and short-answer questions. While these item types have been justified by practice, they are often criticized for lacking face validity (Williamson, Mislevy, and Bejar, 2006, p. 4). Also, discrete-point tests of individual items have been repeatedly characterized as arbitrary and biased because it is practically impossible to design an indirect test that would contain an adequate sample of linguistic phenomena. More importantly, they can only partially reveal learners’ language proficiency. If we think back to the second language acquisition model – “input, apperception, comprehension, intake, integration, output” (Gass, 1997), the last concept in this chain, the *output*, is the real evidence that learners acquired certain linguistic phenomena. Items of indirect measurement can only assess learners’ ability to apperceive and comprehend the input, which may be rather suitable for obtaining information about learners’ receptive language skills such as reading and listening. However, indirect test items are not capable of leading to accurate inferences about learners’ writing and speaking because they do not obtain information on how well

learners integrate the input and how well they can perform in the target language. In order to provide accurate diagnosis of learners' strengths and weaknesses of productive skills, we need to elicit more than recognition; we need to evaluate learners' output, or production.

This implies that computer-based diagnostic assessment needs to expand on its current techniques by adding items that would engage direct measurement of productive skills, i.e., complex constructed-response tasks (Bennett and Ward, 1993). Williamson et al. (2006) emphasize the educational value of such items. Having analyzed a wide range of literature, they argue that constructed responses are beneficial because they:

- “are believed to be more capable of capturing evidence about cognitive processes”
- “provide better evidence of the intended outcomes of educational interventions”
- “offer better evidence for measuring change on both a quantitative [...] and a qualitative level [...]”
- “permit the opportunity to examine the strategies that examinees use to arrive at their solution, which lends itself to collecting and providing diagnostic information” (p. 4).

Considering all these factors, complex constructed responses appear to be most suitable for the purpose of diagnosing constructs from learners' written output in a direct manner. After all, if we want to know about learners' strengths and weaknesses in writing ability, we need to get them to write. Only by observing their extended performance, i.e., how well they can produce texts that are comprehensible, intelligibly organized, register appropriate, correctly punctuated, etc., can we judge their writing proficiency. Moreover, constructed responses based on an adequate task can exhibit various contexts created by learners as well as multiple examples of grammatical structures in use, allowing us to obtain a detailed analysis of their command of grammar. As for vocabulary, these test items would bring diagnosis to the next level by revealing learners' ability to operate with words in order to create comprehensive contexts.

Constructed responses are also advantageous from the viewpoint of practicality. Designing indirect computer-based diagnostic tests as well as any other types of tests requires considerable efforts, especially when it comes to test items. It is very laborious to develop specifications, create a good size pool of items, and pilot the items in order to select the ones that are reliable. In contrast, diagnostic tests based on constructed responses would be more time and cost-efficient in that the test developers would only have to be concerned with elaborating relevant prompts depending on the targeted construct. Further, Alderson (2005) admits that DIALANG “recognized the impossibility of developing specifications for each CEFR-related level separately” (p. 192). This does not seem to be a problem in the case of constructed responses due to the generic nature of prompts. If the same prompt is used by learners of different levels of proficiency, the diagnostic outcome will also be different.

## **5. Automated Scoring Systems**

It is our position that automated writing evaluation is an appropriate, reliable, and efficient way to evaluate constructed responses. Automated scoring, also referred to as computerized essay scoring, computer essay grading, computer-assisted writing

assessment, or machine scoring of essays,<sup>3</sup> today include systems such as AEG (Automated Essay Grading), AES (Automated Essay Scoring), AWE (Automated Writing Evaluation). Despite the numerous terms,<sup>4</sup> these assessment software programs all possess “the ability of computer technology to evaluate and score written prose” (Shermis and Burstein, 2003, p. xiii). The earlier computerized evaluation systems focused on essays, which can be seen in their names, but more recent innovations have expanded the concept of written prose and now include free text or short response answers.

Dikli (2006), Phillips (2007), and Valenti et al. (2003) provide a comprehensive view of existing AES systems, describing their general structure and performance abilities and discussing issues related to their use in testing as well as in the classroom. Here, we will briefly review the most widely used systems in order to further show that their functionality can be extrapolated to diagnostic assessment.

One of the pioneering projects in the area of automated scoring was *Project Essay Grade* (PEG), which was developed in 1966 “to predict the scores that a number of competent human raters would assign to a group of similar essays” (Page 2003, p. 47) It mainly relies on an analysis of surface linguistic features of the text and is designed based on the concepts of *trins* and *proxes*. Trins represent intrinsic variables such as grammar (e.g., parts of speech, sentence structure), fluency (e.g., essay length), diction (e.g., variation in word length), etc., while proxes are the approximations or correlations of those variables, referring to actual counts in student texts. Focusing on writing quality, and relying on the assumption that quality is displayed by the proxes, PEG relies on a statistical approach to generate a score. Recently, PEG has gone through significant modifications, e.g., dictionaries and parsers were acquired, classification schemes were added,<sup>5</sup> and a web-based interface has been employed.

In the late 1990s, the Pearson Knowledge Analysis Technologies produced the *Intelligent Essay Assessor* (IEA) – a set of software tools developed primarily for scoring content related features of expository essays. It is claimed to be suitable for analysis and rating of essays on topics related to science, social studies, history, business, etc. However, it also provides quick customized tutorial feedback on the form related aspects of grammar, style, and mechanics (Landauer, Laham, and Foltz, 2003). Additionally, it has the ability to detect plagiarism and deviant essays. IEA is based on a text analysis method, Latent Semantic Analysis (LSA), and, to a lesser extent, on a number of Natural Language Processing (NLP) methods. This allows the system to score both the quality of conceptual content of traditional essays and of creative narratives (Landauer et al., 2003) as well as the quality of writing. In order to measure the overall quality of an essay, IEA needs to be trained on a collection of domain-representative texts.

The *Electronic Rater* (E-Rater) is a product from the Educational Testing Service in 1997 that has been employed for operational scoring of the Graduate Management Admissions Test Analytical Writing Assessment since 1999. E-Rater produces a holistic score after evaluating the essay’s organization, sentence structure, and content. Burstein (2003) explains that it accomplishes this with the help of a combination of statistical and NLP techniques, which allow for analyses of both content and style. For its model

---

<sup>3</sup> Because this is an emerging field, no single generic terms has been agreed upon yet. The terms are typically used depending on the author’s perspective or on the affiliation of the program described.

<sup>4</sup> We will use AES throughout this paper.

<sup>5</sup> These additions were explored with a number of tests to find whether they could indicate distinctions among levels of writing proficiency.

building, E-Rater uses a corpus-based approach,<sup>6</sup> the corpora containing unedited first-draft essay writing. Outputs for model building and scoring are provided by several independent modules. The syntactic module is based on a parser that captures syntactic complexity; the discourse module analyzes the discourse-based relationship and organization with the help of cue words, terms, and syntactic structures; and the topical analysis module identifies the vocabulary use and topical content.

The *Bayesian Essay Test Scoring System* (BETSY), funded by the Department of Education and developed at the University of Maryland, was also designed for the purpose of automated scoring. BETSY relies on a statistical technique based on a text classification approach that, as Valenti et al. (2003) claim, may combine the best features of PEG, LSA, and E-Rater. A large set of essay features are analyzed, among which there are content-related features (e.g., specific words and phrases, frequency of content words) and form-related features (e.g., number of words, number of certain parts of speech, sentence length, number of punctuation marks). Rudner and Liang (2002) assert that this system can also be used in the case of short essays, can be applied to various content areas, can be employed to provide a classification on multiple skills, and, finally, can allow for obtaining diagnostic feedback in addition to scoring.

A product of Vantage Learning used for the rating of the Analytical Writing Assessment section of the GMAT is *IntelliMetric*. It is the first automated scoring system that was developed on the basis of artificial intelligence (AI) blended with NLP and statistical technologies. IntelliMetric is “a learning engine that internalizes the characteristics of the score scale [derived from a trained set of scored responses] through an iterative learning process”, creating a “unique solution for each stimulus or prompt” (Elliot, 2003, p. 71). To attain a final score, more than 300 semantic, syntactic, and discourse level features are analyzed by this system. They can be categorized into five groups: focus and unity (i.e., cohesiveness and consistency in purpose and main idea), development and elaboration (i.e., content through vocabulary use and conceptual support), organization and structure (i.e., logical development, transitional flow, relationship among parts of the response), sentence structure (i.e., syntactic complexity and variety), and mechanics and conventions (i.e., punctuation, sentence completeness, spelling, capitalization, etc.). Apart from the scoring ability, IntelliMetric’s modes allow for student revision and editing as well as for diagnostic feedback on rhetorical, analytical, and sentence-level dimensions.

In 1999, the *Automark* software system was developed in the UK as an effort to design robust computerized marking of responses to open-ended prompts. The system utilizes NLP techniques “to perform an intelligent search of free-text responses for predefined computerized mark scheme answers” (Mitchell, Russel, Broomhead, and Aldridge, 2002, pp. 235-236). Automark analyzes the specific content of the responses, employing a mark scheme that indicates acceptable and unacceptable answers for each question. The scoring process is carried out by a number of modules: syntactic preprocessing, sentence analysis, pattern matching, and feedback. The latter is provided as a mark, but more specific feedback is also possible (Valenti, 2003). What makes it similar to human raters is the fact that, while assessing style and content, it can ignore errors in spelling, typing, syntax, and semantics that do not interfere with comprehension.

---

<sup>6</sup> A corpus-based approach is different from a theoretical approach in which features are hypothesized based on characteristics expected to be found in the data sample.

To analyze the constructed input and to produce scores and feedback, each of the systems described above uses various techniques – statistical, NLP. The following section will describe these techniques and explain how they can be applied for the diagnosis of various components of writing proficiency.

## 6. Techniques and constructs

*Statistical techniques* of text features typically require an initial training phase for parameter estimation. There are several types of statistical analyses incorporated into the automated scoring systems considered above. For example, E-Rater employs “simple keyword analysis,” which looks for coincident keywords between the student essay and the scored one. PEG relies on “surface linguistic features analysis” that finds the features to be measured and uses them as independent variables in a linear regression to yield the score. IEA, in turn, is underpinned by “latent semantic analysis (LSA),” a complex statistical technique developed for information retrieval and document indexation (Deerwester, Dumais, Landauer, Furnas, and Harshman, 1990). LSA finds repeated patterns in the student response and the reference text to extract the conceptual similarity between them. Finally, BETSY is based on “text categorization” techniques, which can consist of several score categories, associate the student response with one of them, and assign the score accordingly.

*Natural Language Processing techniques* apply computational methods for the analysis of natural language (Burstein, 2003). They include syntactic parsers or rhetorical parsers, the former being able to evaluate the linguistic structure of a text, and the latter the discourse structure. Combining these techniques can enhance the use of statistics by involving deep-level parsing and semantic analysis, therefore gathering more accurate information about the student’s response and providing a more accurate assessment. Among the current scoring systems, E-Rater, Automark,<sup>7</sup> and IntelliMetric successfully employ NLP. IntelliMetric, in addition to NLP, exploits AI techniques. Dikli (2006) claims that this system is “modeled on the human brain,” being based on a “neurosynthetic approach [...] used to duplicate the mental processes employed by the human expert raters” (p. 17). Unlike the statistical procedures, NLP is very difficult to implement across languages.

Recent empirical work provides evidence that E-Rater, IEA, PEG, IntelliMetric, Automark, and BETSY are valid and reliable (Burstein, 2003; Elliot, 2003; Keith, 2003; Landauer et al., 2003; Mitchell et al., 2002; Page, 2003; Valenti et al., 2003). The main method employed for system validation is single essay agreement results with human ratings. Summarizing research results, Dikli (2006) concludes that correlations and agreement rates between the system and human assessors are typically high. Experiments on PEG obtained a multi-regression correlation of 87%. E-Rater has scored essays with agreement rates between human raters and the system consistently above 97%. BETSY achieved an accuracy of over 80%. Automark’s correlation ranged between 93% and 96%. IEA yielded a percentage for an adjacent<sup>8</sup> agreement with human graders between

---

<sup>7</sup> Automark also makes use of an information extraction approach, which is considered a shallow NLP technique as it typically does not require a full-scale analysis of texts.

<sup>8</sup> Adjacent agreement is different from exact agreement in that it requires two or more raters to assign a score within one scale point of each other (Elliot, 2003).

85% and 91%. IntelliMetric also reached high adjacent agreement (98%), and the correlation for essays not written in English attained 84%.

With such functionality, the potential of scoring systems for diagnostic assessment is undeniably apparent. As mentioned before, to diagnose students' writing proficiency automatically, we would need to be guided by the construct<sup>9</sup> – what we want to diagnose. Similar to DIALANG, we would want our automated diagnosis program to be able to analyze and evaluate students' knowledge of writing in terms of textual organization, register appropriacy, grammatical accuracy, and vocabulary. Alderson (2005) admits that tests of such detailed knowledge aspects are “difficult to construct because of the need to cover a range of contexts and linguistic backgrounds, and to meet the demands of reliability” (p. 257). Automated scoring systems can, in principle, assess these, plus other construct components (see Table 2); moreover, they can do that with more precision and objectivity due to the sophistication of their techniques. Even more importantly, the inferences based on their results can be more valuable because such systems can uncover aspects of the productive writing proficiency from students' constructed responses, which existing CALT diagnostic tests cannot achieve with their indirect measures limited to recognition.

Table 2. Techniques used in Automated scoring systems.

<b>System</b>	<b>Constructs</b>	<b>Technique</b>
PEG (Page, 2003)	Grammar, fluency, diction	Statistical (measurement of surface linguistic features)
IEA (Landauer et al., 2003)	Content Grammar, style, mechanics Plagiarism and deviance	Statistical (Latent Semantic Analysis)
E-Rater (Burstein, 2003)	Topical content Rhetorical structure Syntactic complexity	Statistical (e.g., vector analyses) Natural Language Processing (NLP) (e.g., part-of-speech taggers)
BETSY (Rudner and Liang, 2002)	Content Grammar, style, mechanics	Statistical (Bayesian text classification)
IntelliMetric (Elliot, 2003)	Focus / unity Development / elaboration Organization / structure Sentence structure Mechanics / conventions	Artificial Intelligence (AI) Natural Language Processing (NLP) Statistical
Automark (Mitchell et al., 2002)	Content Grammar, style, mechanics	Natural Language Processing (NLP)

It must be acknowledged, however, that all these systems were designed for the evaluation of native-speaker writing. While there is no doubt that their ability to analyze free production would be extremely valuable in assessing non-native speaker responses, there might be questions as to whether such systems can be as reliable in the case of ESL/EFL. Indeed, computerized assessment of constructed responses produced by non-

<sup>9</sup> Scoring levels are also essential, but the focus of this paper does not include issues pertaining to proficiency levels.

native speakers, especially at low levels of proficiency, is prone to face barriers in dealing with ill-formed utterances. Research in this area is only at its outset; however, recent implementations and insights seem to be encouraging. For instance, in practical terms, Educational Testing Service has been successfully employing E-Rater to evaluate ESL/EFL performance on the TOEFL exam. Research-wise, Burstein and Chodorow (1999) found that the features considered by E-Rater are generalizable from native speaker writing to non-native speaker writing and that the system was not confounded by non-standard English structures. Leacock and Chodorow (2003) also claim that recent advances in automatic detection of grammatical errors are quite promising for learner scoring and diagnosis. In line with this idea, Lonsdale and Strong-Krause (2003), having explored the use of NLP for scoring novice-mid to intermediate-high ESL essays, claim that “with a robust enough parser, reasonable results can be obtained, even for highly ungrammatical text” (p. 66). Undoubtedly, much improvement is needed to construct automated scoring systems that would capture the distinctiveness of learner language, but this can be achieved by integrating a combination of scoring techniques, which will allow for building diagnostic models of learner writing.

## **7. Advantages of automated writing tests**

Due to the auspicious potential of these systems, we have reasons to believe that automated scoring would be a promising innovation for diagnostic writing assessment. A major advantage would be the ability to automatically evaluate “qualities of performance or work products” (Williamson et al., 2006) by analyzing evidence that would allow for making inferences about strengths and weaknesses in learner’s writing proficiency. With such evidence, automated evaluation would not be a mere application of new technologies, but it would become an essential component of the validity argument for the use of automated diagnostic tests. Moreover, the focus on evidentiary reasoning would facilitate the development of automated diagnostic tests if we choose to follow the framework of Evidence-Centered Design, which “is an approach to constructing and implementing educational assessments in terms of evidentiary arguments” (Mislevy, Steinberg, Almond, and Lukas, 2006, p. 15).

Further, considering that an inherent characteristic of diagnostic assessment is its orientation towards learning, the provision of meaningful<sup>10</sup> feedback would be an essential strength of automated diagnostic writing tests. They could potentially provide complex<sup>11</sup> evaluative responses (see Table 3) based on positive empirical evidence, making it possible for learners to take steps towards remediation and improvement. An automated test could also enhance the learning opportunity by allowing learners to act upon the received feedback, re-submit their texts, and make gradual improvements in their self-instruction efforts. Moreover, because the systems are generally trained on certain material, directed feedback could be linked to the training texts (Landauer et al., 2003) (which could be either authentic or learner texts), thus making diagnostic assessment more interactive, tailored to both instruction and to individual learners. For examples of systems that have already implemented tools which enhance their summative

---

<sup>10</sup> Heift (2003) defines meaningful feedback as a “response that provides a learning opportunity for students.” (p. 533)

<sup>11</sup> The automated feedback could possess many or even all the characteristics presented in Table 3 since they are not necessarily mutually exclusive.

evaluation function with instructional applications, interested readers can look into *Criterion<sup>SM</sup>* by Educational Testing Service or *MY Access* by Vantage Learning.

Table 3. Feedback leading to better learning

<i>Positive evidence</i>
1. Explicit feedback (Caroll, 2001; Caroll and Swain, 1993; Ellis, 1994; Lyster, 1998, Muranoi, 2000)
2. Individual specific (Hyland, 1998)
3. Metalinguistic feedback (Rosa and Leow, 2004)
4. Negative cognitive feedback (Ellis, 1994; Long, 1996; Mitchell and Myles, 1998)
5. Intelligent feedback (Nagata, 1993, 1995)
6. Output-focused feedback (Nagata, 1998)
7. Detailed iterative feedback (Hyland and Hyland, 2006)
8. Feedback – accurate, short, one at a time (Van der Linden, 1993)

In addition to being more dependable in assessing writing skills and providing diagnostic feedback, automated writing tests would have a number of other advantages. Dikli (2006) emphasizes that scoring systems can enhance practicality, helping overcome time, cost, reliability, and generalizability issues. Assessment of writing implies design of prompts, creation of rubrics, training of raters, and human scoring the responses. Indisputably, automated scoring can reduce the need for some of these activities. What can even be eliminated is the need to incorporate an initial placement component, like the one DIALANG has, since the system's analyses of constructed responses can both describe learners' performance and place them in the appropriate level. In fact, because learners' written production can be analyzed in so many possible details, one can argue that there is no need for designing separate tests for individual skills such as grammar, vocabulary, etc.

On a larger scale, essay grading is criticized for "perceived subjectivity of the grading process" (Valenti et al., 2003, p. 319) because of the frequent variation in scores assigned by different raters. Automated evaluation can increase objectivity of assessment, providing consistency in scoring and feedback through greater precision of measures (Phillips, 2007). Also, the systems, if re-trained, can re-score student answers should the evaluation rubric be redefined (Rudner and Gagne, 2001). Finally, automated diagnostic tests can have built-in validity checks to address possible biases (Page, 2003).

## 8. Conclusion

With regard to diagnosis and learning, there is very little work despite the technical capabilities currently available (Chapelle, 2006). In the past few years, several successful automated scoring systems have been developed; their use is rapidly growing, which can and should positively affect the field of CALT. Due to the ability to describe results and provide individualized feedback in a variety of ways, automated evaluation can generate a new wave of research on computerized diagnostic testing, thus increasing our understanding of the nature of language proficiency and of the way learners develop from level to level. This can also be achieved through the insights gained from learner corpora

used in automated systems for training purposes. In the long run, such an understanding could contribute to the formulation of a more specific writing proficiency framework, and, therefore, of a better theory of diagnosis. Other issues regarding automated diagnosis of written language proficiency that would require future empirical investigations are design, validation, and evaluation of diagnostic tests. Of equal value would be studies on the effectiveness of automated feedback as well as on the impact of automated diagnostic writing tests on instruction remediation. These suggestions for future research are far from being exhaustive, and we encourage the readers to think of further questions since “the potential of automated essay evaluation [...] is an empirical question, and virtually no peer-reviewed research has yet been published that examines students’ use of these programs or the outcomes.” (Hyland and Hyland, 2006, p. 109)

To conclude, because diagnostic tests “should be thorough and in-depth, involving an extensive examination of variables” (Alderson, 2005, p. 258), it is very important for the profession to consider automated evaluation as an exceptional possibility for innovating computerized diagnostic test methods. Diagnostic writing tests, in particular, need to develop from computer-based indirect measures of recognition to automated systems-based direct assessment of written language production and proficiency. We have attempted to justify our argument by pointing out the latent advantages of automated analysis of constructed responses and of automated feedback for developing learners’ writing proficiency. However, we acknowledge that this venture is not an easy one. Designing an automated diagnostic writing test that satisfies all the necessary constraints will require a lot of incremental work. Only through close collaboration with computer scientists, computational linguists, specialists in SLA, CALL, and other related areas, can we achieve desired results.

## References

- Alderson, J.C. (2005). *Diagnosing foreign language proficiency: the interface between learning and assessment*. London: Continuum.
- Alderson, J.C., Clapham, C.M., and Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.
- Alderson, J.C. and Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing* 22, pp. 301-320.
- American Council for the Teaching of Foreign Languages (1983). *ACTFL proficiency guidelines* (revised 1985). Hastings-on-Hudson, NY: ACTFL Materials Center.
- Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L.F. and Palmer, A.S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bejar, I. (1984). Educational diagnostic assessment. *Journal of Educational Measurement*, 21(2), pp. 175-189.
- Bennet, R., and Ward, W. (Eds.) (1993). *Construction versus choice in cognitive measurement: Issues in constructed responses, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Burstein, J. (2003). The E-rater scoring Engine: Automated Essay Scoring with Natural Language Processing. In Shermis, M.D. and Burstein, J.C. *Automated essay*

- scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Burstein, J., and Chodorow, M. (1999). Automated essay scoring for nonnative English speakers. In *Proceedings of the ACL99 workshop on computer-mediated language assessment and evaluation of natural language processing*. College Park, MD. Retrieved on August 3, 2007 at [http://www.ets.org/Media/Research/pdf/erater\\_acl99rev.pdf](http://www.ets.org/Media/Research/pdf/erater_acl99rev.pdf)
- Burstein, J., Chodorow, M., and Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine*, 25(3), 27-36.
- Carroll, S. (2001). *Input and evidence: The raw material of second language acquisition*. Amsterdam: Benjamins.
- Carroll, S., and Swain, M. (1993). Explicit and implicit negative feedback: An empirical study of the learning of linguistic generalizations. *Studies in Second Language Acquisition*, 15, pp. 357–366.
- Chapelle, C. (2006). Test Review. *Language Testing*, 23, pp. 544-550.
- Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, and assessment*. Cambridge: Cambridge University Press.
- Cowie, J. and Lehnert, W. (1996). Information Extraction. *Communications of the ACM*, 39(1), pp. 80-91.
- Darus, S. and Stapa, S. (2001). Lecturers' expectations of a computer-based essay marking system. *Journal of the Malaysian English Language Teaching Association*, XXX, pp. 47-56.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and McNamara, T. (1999). *Dictionary of language testing*. Cambridge: University of Cambridge Local Examination Syndicate and Cambridge University Press.
- Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6), pp. 391-407.
- Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, 5(1), p.4.
- Elliot, S. (2003). IntelliMetric: From here to validity. In Shermis, M.D. and Burstein, J.C. *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Ellis, R. (1994). A theory of instructed second language acquisition. In N. C. Ellis (Ed.), *Implicit and explicit learning of languages* (pp. 79–114). San Diego, CA: Academic Press.
- Gass, S. (1997). *Input, interaction, and the second language learner*. Mahwah, NJ: Erlbaum.
- Hughes, A. (1989). *Testing for language teachers*. 2<sup>nd</sup> Ed. Cambridge: Cambridge University Press.
- Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing*, 7(3), pp. 255-286.
- Hyland, K., and Hyland, F. (Eds.). (2006). *Feedback in Second Language Writing: Contexts and issues*. New York: Cambridge UP.
- Keith, T. (2003). Validity of automated essay scoring systems. In Shermis, M.D. and Burstein, J.C. *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Landauer, T., Laham, D., and Foltz, P. (2003). Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. In Shermis, M.D. and Burstein, J.C. *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Leacock, C. and Chodorow, M. (2003). Automated grammatical error detection. In Shermis, M.D. and Burstein, J.C. *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie and T. K. Bhatia (Eds.), *Handbook of second language acquisition*. San Diego, CA: Academic Press.
- Lonsdale, D, and Strong-Krause, D. (2003). Automated rating of ESL essays. *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing*, 2, pp. 61 – 67. Retrieved on July 20, 2007 at <http://ucrel.lancs.ac.uk/acl/W/W03/W03-0209.pdf>.
- Lyster, R. (1998). Negotiation of form, recasts, and explicit correction in relation to error types and learner repair in immersion classrooms. *Language Learning*, 48, pp. 183–218.
- Mislevy, R., Steinberg, I., Almond, R., and Lukas, J. (2006). Concepts, terminology, and basic models of Evidence-Centered Design. In Williamson, D., Mislevy, R., and Bejar, I. (Eds.) (2006). *Automated scoring of complex tasks in computer-based testing* (pp. 15-47). Mahwah, NJ: Lawrence Erlbaum Associates.
- Mitchell, R., and Myles, F. (1998). *Second language learning theories*. London: Arnold Publishers.
- Mitchell, T., Russel, T., Broomhead, P., and Aldridge, N. (2002). Towards Robust Computerised Marking of Free-Text Responses. In: *Proceedings of the 6th CAA Conference*, Loughborough: Loughborough University. Retrieved on July 31, 2007 at [https://dspace.lboro.ac.uk/dspace/bitstream/2134/1884/1/Mitchell\\_t1.pdf](https://dspace.lboro.ac.uk/dspace/bitstream/2134/1884/1/Mitchell_t1.pdf).
- Moussavi, S., A. (2002). *An encyclopedic dictionary of language testing*. 3<sup>rd</sup> Ed. Taiwan: Tung Hua Book Company.
- Muranoi, H. (2000). Focus on form through interaction enhancement: Integrating formal instruction into a communicative task in EFL classrooms. *Language Learning*, 50, pp. 617–673.
- Nagata, N. (1998). The relative effectiveness of production and comprehension practice in second language acquisition. *Computer Assisted Language Learning* 11 (2), pp. 153–77.
- Nagata, N. (1995). An effective application of natural language processing in second language instruction. *CALICO Journal* 13(1), pp. 47–67.
- Nagata, N. (1993). Intelligent computer feedback for second language instruction. *The Modern Language Journal* 77 (3), pp. 330–9.
- Page, E. (2003). Project Essay Grade. In Shermis, M.D. and Burstein, J.C. *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Pawlikowska-Smith, G. (2000). *Canadian language benchmarks 2000: English as a second language for adults*. Ottawa: Citizenship and Immigration Canada.

- Phillips, S. (2007). *Automated essay scoring: A literature review*. Society for the Advancement of Excellence in Education. Retrieved July 12, 2007 from <http://www.saeec.ca/pdfs/036.pdf>.
- Rosa, E., and Leow, R. (2004). Awareness, different learning conditions, and second language development. *Applied Psycholinguistics*, 25, pp. 269–292.
- Rudner, L. and Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation*, 7(26). Retrieved July 23, 2007 from [http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content\\_storage\\_01/0000019b/80/19/5e/46.pdf](http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/19/5e/46.pdf).
- Rudner, L. and Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2), pp. 3-21.
- Scalise, K. and Bernard, G. (2006). Computer-based assessment in e-learning: A framework for constructing “intermediate constraint” questions and tasks for technology platforms. *Journal of Technology, Learning and Assessment*, 4(6), pp. 4-43.
- Schonell, F., J. and Schonell, F. E. (1960). *Diagnostic and attainment testing, including a manual of tests, their nature, use, recording, and interpretation*. Edinburgh: Oliver and Boyd.
- Shermis, M.D. and Burstein, J.C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Valenti, S, Nitko, A., and Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education (Information Technology for Assessing Student Learning Special Issue)*, 2, pp. 319-329.
- Van der Linden, E. (1993). Does feedback enhance computer-assisted language learning. *Computers & Education*, 21 (1-2), pp. 61-65.
- Williamson, D., Mislevy, R., and Bejar, I. (Eds.) (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Wylie, E. and Ingram, D. (1995/99). *International Second Language Proficiency Ratings (ISLPR): General proficiency version for English*. Brisbane: Center for applied Linguistics and Languages, Griffing University.