*Assessment in Education, Vol. 10, No. 3, November 2003*

Carfax Publishing
Taylor & Francis Group

# Automatic Essay Assessment

THOMAS K. LANDAUER
*University of Colorado and Knowledge Analysis Technologies, USA*

DARRELL LAHAM
*Knowledge Analysis Technologies, USA*

PETER FOLTZ
*New Mexico State University and Knowledge Analysis Technologies, USA*

ABSTRACT *Computational techniques for scoring essays have recently come into use. Their bases and development methods raise both old and new measurement issues. However, coming principally from computer and cognitive sciences, they have received little attention from the educational measurement community. We briefly survey the state of the technology, then describe one such system, the Intelligent Essay Assessor (IEA). IEA is based largely on Latent Semantic Analysis (LSA), a machine-learning model that induces the semantic similarity of words and passages by analysis of large bodies of domain-relevant text. IEA's dominant variables are computed from comparisons with pre-scored essays of highly similar content as measured by LSA. Over many validation studies with a wide variety of topics and test-takers, IEA correlated with human graders as well as they correlated with each other. The technique also supports other educational applications. Critical measurement questions are posed and discussed.*

While most educators believe that good assessment is crucial to education, many worry about insufficient correspondence between measured and desired abilities. As a result there are growing efforts to develop more ecologically valid methods, including constructed response and performance-based tests. The particular problem with which we are concerned is that current methods appear to encourage teaching strategies that provide too little practice in formulating knowledge in self-generated prose. One obvious response is more essay-style exams. However, there are serious problems with essay assessments. They are too expensive in time and money to be used as often as would be valuable, and they often have low reliability.

A potential solution to this dilemma is the development of pedagogically adequate, psychometrically sound, and socially acceptable machine assessment and tutorial feedback for expository essays (Clauser *et al.*, 1997; Clyman *et al.*, 1995). Methods for automatic scoring of test essays are just now coming into use. Most of them are based on artificial intelligence, computational linguistics, and cognitive

ISSN 0969-594X print; ISSN 1465-329X online/03/030295-14 ©2003 Taylor & Francis Ltd
DOI: 10.1080/0969594032000148154

science foundations, rather than on traditional psychometrics and educational measurement. Therefore we think it is important to bring them to the critical attention of the educational measurement community. We begin with a brief review of the state of the art in machine scoring and analysis of essays. Then, as a concrete example, we describe theoretical and technical bases and evaluative evidence concerning key aspects of the particular method that we ourselves have developed, giving special attention to measurement issues.

## The State of the Art in Computer Analysis and Scoring of Essays

First a caveat. Providers of currently available systems have not revealed all of the potentially important details of their methods. This may be an understandable state of affairs in today's highly competitive software technology market, but it creates difficulties for public confidence and professional debate. We hope to ameliorate the situation here by giving a description of our own system that is sufficiently detailed to pose some of the psychometric problems involved. However, we cannot reveal more about the other systems than has been made public, and our review must, therefore, be short and somewhat sketchy. The general technical approach to automated essay marking has been first to identify a large pool of computationally measurable characteristics of machine-readable test essays that are potential correlates of human scores. Subsets of the initial pool are then selected and combined by familiar statistical methods, such as multiple regression, against a collection of humanly scored essays (Burstein, 2003; Landauer *et al.*, 2003; Page, 1966, 1994). The variables have been taken from many sources: components of readability measures such as sentence length and word frequency distributions; components of computerised grammar and spelling analysis programs such as detection of passive and run-on sentences; variables from computational linguistic models for discourse understanding and generation, such as trigger terms diagnostic of argument structure; measures used in information retrieval such as keyword matching; variables used in automatic speech modelling, such as word bigram frequencies; and easily computed surface variables such as length in words, excessive word repetition, and words not found in an appropriate dictionary. Other variables are novel detectors of discourse or literary qualities implemented as artificial intelligence style rules, simple counts, or new statistical analyses. For example, our system adds a new corpus-based computational model of word and passage meanings to select pre-scored examples for comparison. Others may add new techniques of which we are unaware.

The harvest-and-winnow approach resembles the manner in which traditional test items are assembled. The variables are often considered index variables or 'proxies' defended primarily by their empirical correlations with overall human scores. The feature pool sizes vary over a wide range, apparently from around ten to over 100. The methods of selection also probably differ. Of the individual variables reported some have clear face validity; some linguistic, psycholinguistic, computational linguistic, psychological, literary, or artificial intelligence theory justifications. Few have demonstrated external criterion validity.

Overall validation, where reported, has been almost exclusively performed by correlation or score agreement between system and independently assigned human scores. Validity is assessed by comparing the correspondence between the system's score and human scores with that between two or more independent human scores. Often comparisons are made by counting the proportion of discrete scores on, say, a six-point scale, that are the same by machine as by human scoring, or within one scoring point. While these measures appear to have pragmatic value in sharing results with non-professional audiences, they confound the accuracy of the machine analysis of essay quality, which may yield a more refined or continuous score, with the logically separate step of matching monotone-discrete or nominal human scoring category assignments.

As far as the number of essays available for training, the amount of essay-specific tuning involved, the kind of essay—argument, opinion, knowledge exposition, creative narrative, etc.—and other variables allow comparison, all the methods appear to have nearly the same general level of accuracy. They are reported to agree with human readers 'about', 'nearly', 'as', or 'at least as' well as human readers agree with each other. Some are reported to agree with the average of multiple readers better than with a single reader. It appears that all do better the more pre-scored essays they have for training data. There have been no systematic comparisons between systems with respect to their relative accuracy as a function of size of training sets for the same essays. We have done studies showing that IEA accuracy asymptotes at from 100–400 pre-scored essays depending on the topic and prompt.

The overall accuracy results raise some of the educational and social issues mentioned earlier. Reliabilities of essays *per se* are below those of other kinds of tests. To our knowledge, repeatability has never been empirically determined. The fact that the methods all depend on correspondence with intuitive human ratings suggests important but complex aspects of the problem, some of which will be elaborated below.

## The Intelligent Essay Assessor (IEA)

From this point on we couch our discussion in terms of the system we have developed, the *Intelligent Essay Assessor.* Like other systems IEA uses a statistical combination of several measures to produce an overall score. It differs from others in their number and nature, focuses more on topical content rather than writing quality, is aimed at tutorial feedback in addition to summative assessment, and relies primarily on corpus statistical measures rather than ones motivated by traditional artificial intelligence (AI) or linguistic theory. Its greatest difference lies in its use of Latent Semantic Analysis (LSA), a machine-learning model of human understanding of text, on which its training and calibration methods, and the ways it is used tutorially, depend. LSA is mathematically complex, and a full description is beyond the scope of the present paper—details can be found elsewhere (see Berry, 1992; Berry *et al.*, 1995; Deerwester *et al.*, 1990; Eckart & Young, 1936; Golub & Van Loan, 1989; Landauer & Dumais, 1997; Landauer *et al.*, 1998). While the LSA model of verbal meaning at first appears to be an implausible over-simplification, it turns out to yield remarkably accurate simulations of a wide spectrum of language phenomena, and robust support for automation of many language-dependent tasks. Extensive psychological and mathematical rationales and empirical evidence for LSA can be found in the references cited above (especially Deerwester *et al.*, 1990; Foltz, 1996; Kintsch *et al.*, 2000; Laham, 1997, 2000; Landauer & Dumais, 1997; Landauer *et al.*, 1998; Rehder *et al.*, 1998).

LSA does not always give intuitive results on relations between phrases or sentences, especially where local syntactic influences are strong, but it usually does well with paragraphs or 50- to 300-word essay-like passages. LSA has produced close approximations to the similarity to humans of the verbal meaning of words and passages as exhibited in a considerable variety of well-known verbal phenomena and scientific and practical applications. Here are some examples.

LSA was tested on 80 retired multiple-choice vocabulary items from the Educational Testing Service Test of English as a Second Language (TESL). LSA was correct on 60 of the 80 items, matching the average of a large sample of students from non-English-speaking countries who had applied for admission to US colleges (Landauer & Dumais, 1997).

In a second set of simulations, LSA was trained on a popular introductory psychology textbook and tested with the same multiple-choice final exams as students in two large classes. LSA's score was lower than the class averages but passing in both. In both cases its errors resembled those of students (Landauer *et al.*, 1998).

There are two characteristics of LSA that make it well suited for IEA purposes: first, it accurately measures the similarity of conceptual content of two essays despite the possible use of differing vocabulary; second, although in other contexts LSA's neglect of word order can be a flaw, here it is also a virtue. This is because it chooses essays for comparison without direct regard to factors such as grammatical and syntactic correctness, discourse flow, and poetics, whose assessments we believe to be better left to separate measurement. Nevertheless the content score assesses more than knowledge since human grading judgements are influenced by other factors. And the LSA representation, being based on the complete combination of words, carries information about expressive factors as well. One especially bothersome critical response was adopted by several newspaper editorialists and other motivated sceptics in times before we added syntactic measures and validity checks. They first wrote a good essay, then scrambled its words and sentences, or sprinkled it with 'nots', and observed that IEA still gave a good score. Explaining that IEA would not give a good score unless the original were good, and that it is virtually impossible to write a good essay in scrambled order or with systematically wrong information in the first place, and that no student would have any reason to try these tricks, was rarely of any avail.

## How IEA Works

To compute a total score, IEA combines three kinds of variables, which we refer to as *content*, *style* and *mechanics*, plus *validity* and *confidence* measures. The architecture is illustrated in Figure 1.

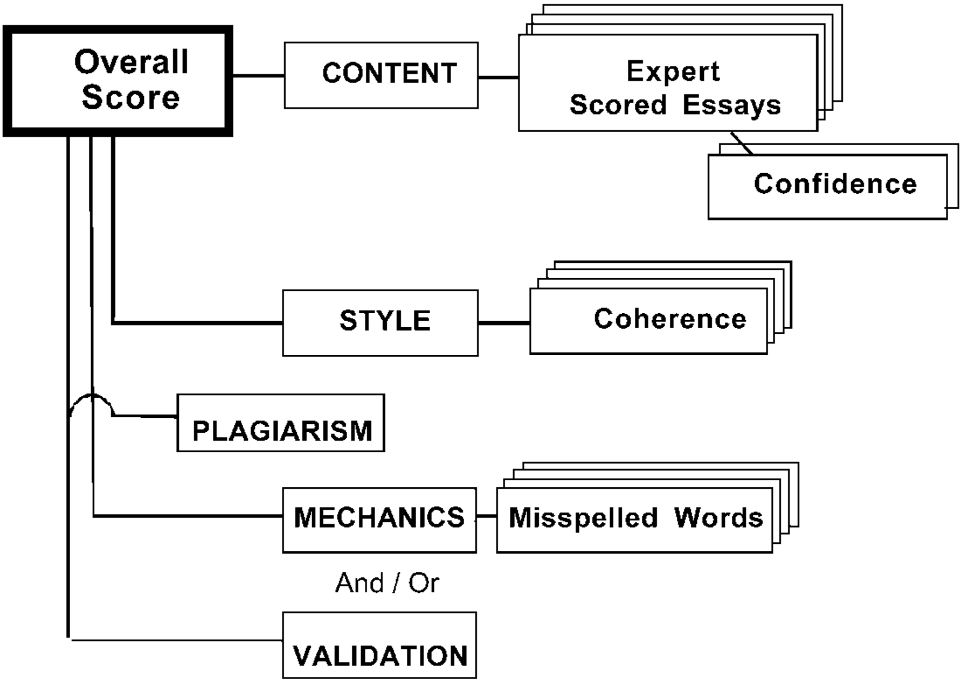

FIG. 1. Architecture for Intelligent Essay Assessor scoring.

Default use combines the components by a form of constrained multiple regression on human scores in a training sample. One constraint is that the content score is always given the greatest weight. Internal measures of scoring confidence, and triggers for highly unusual content, plagiarism, and other forms of counterfeiting are always computed and used to flag essays for human examination. It is not possible to describe every detail of the mechanisms used in IEA here, but to offer an insight into the workings of IEA, the content component will be described in some detail.

## The IEA Content Component

The content variable is based on what we call a *direct prediction* of the score that expert humans would give to an essay. What we mean by a direct prediction is one whose independent variables are natural human judgements of student essays, parts or aspects thereof, not theoretical or empirical index variables or proxies. Here, in stepwise fashion, is the procedure.

1. Apply LSA to extensive background text to represent the meaning of words as used in the domain of a test. IEA does a better job the more relevant background text it gets, for example, a whole biology or psychology textbook, rather than just a section or chapter, or just the essays themselves (Rehder *et al.*, 1998). Training on just the essays (as is necessarily the case in standard information retrieval approaches), although adequate with large numbers of essays, produces consistently inferior results.

2. Have a representative sample of essays scored by human readers.
3. Represent every training and to-be-scored essay as an LSA vector.
4. Compute similarity between each to-be-scored essay and each of the previously scored essays.
5. Choose a certain number, *k* of pre-scored essays with which to compare the to-be-scored essay. Notice that in choosing just *k* comparison essays for the prediction IEA is actually using the relation of the to-be-scored answer to *all* the pre-scored essays because it is deciding which scores are to be considered irrelevant. This is important because in many cases we want it to be possible that two students could write very different answers, both of which are good.
6. Compute a prediction of the grade that the readers of the comparison essays would have given the to-be-scored essay. Intuitively, this may be thought of as a vicarious human judgement, one based on a number of highly similar experiences.

Unlike other methods, direct prediction does not first try to predict intuitive judgements from a set of index variables, then use the index variables in its place. The latter method could only be as good or better than direct prediction, as we define it, if it could predict human scores more accurately than human scores predict each other. This is not impossible. For example, tread thickness measurement of tyres would probably predict failure in the second 30,000 miles of use better than the experience rate of other tyres. The doubled error variance of using inconsistent human judgements as predictors of other inconsistent human judgements, or error in the manner in which one is used to predict the other (e.g. in IEA by its relation of one essay to another through LSA similarity measures) may leave room for such a possibility. However, it seems unlikely that better variables have been identified for essay judgements. While awaiting such a discovery it seems prudent to (in part) respect the widespread folk opinion that only expert humans are fit judges of essays. If nothing else, so doing may make the technology more publicly acceptable.

This technique has another important property. An essay exam answer may be written in a variety of good ways. In addition to wording the same content in different ways, the actual content can sometimes be different but equally good (or bad). This is especially true, of course, in creative narrative, argument, or opinion papers. However, something similar is also possible in substantive topics. Consider, for example, the question: 'Discuss important factors in the battle of Gettysburg'. One student may dwell on the military action in terms of troops and charges, another on strategy and information issues, another on the politics of command. Perhaps a really good answer might manage to do a good job on all three aspects. But an adequate answer might well focus on just one or two. Moreover, different criteria may apply to different kinds of answers. Knowing who led crucial charges at Gettysburg is important only for essays focusing on the battle itself; the name Pickett is extraneous in an essay stressing errors of logistics and information. Thus, the context specificity of the direct prediction procedure approximately implements the goals of item response theory for essay scoring, although the different treatment of 'items' in different performance contexts is obviously not achieved in the same way. Direct prediction restores much of the expert ad hoc, non-monotonic reasoning in evaluation whose loss alarms some critics of automatic scoring methods. On the other hand, because the method requires consensus over multiple comparisons, it also tames the erratic idiosyncrasies possible in unconstrained intuition.

IEA's capability of dealing with unusual content is surely not perfect; the essays in the comparison set are unlikely to be perfect predictors, and the algorithm for extrapolating from their scores to the unknown essay cannot be error-free. Safeguards to reduce the number of such errors are possible. If there are too few sufficiently similar comparison essays, or the comparison scores are not sufficiently in agreement, IEA flags the essay for examination.

An additional useful characteristic of the direct prediction approach is that it can be easily applied to many properties of an essay other than its overall quality. All that is required is that human readers make judgements of a particular property according to a common notation or scale—for example, how persuasive the essay is as a whole.

## The Reliability and Validity of Automatic Essay Scoring

In the sense of agreement with human judges, the validity of essay grades is identical to their reliability. However, there are other possible criteria, such as correlation with other measures of the same knowledge, or better correlation with more expert than with less expert judges. We will touch on such measures later. First we report studies of the reliability of IEA versus human graders as compared to the reliability between human graders. For validation, we prefer to deal with product-moment correlations between continuous IEA scores and whatever score the human graders give, rather than the common practice of measuring agreement on grade or score categories. This reduces to the minimum the contribution of quantification error [1]. It also avoids classifying the scores into discrete score groups, a matter that involves pedagogical and political decisions largely irrelevant to the questions of validity.

In each case, we first collected a large number of essays written to the same prompt by students in a real test administration. These were provided either by large national or international professional testing organisations such as Educational Testing Service (ETS) or the Center for Research on Evaluation, Standards, and Student Testing (CRESST) or by professors at major universities. The essays in each case were graded independently by at least two graders who were knowledgeable in the content domain and quality criteria of the test and trained in its scoring according to either holistic rubrics or analytic components. Graders were always blind to the IEA scores. In the case of professional scoring and most educational institution scoring, readers were not told that an automatic scoring system would be used, and were unaware of the existence or underlying technology of IEA. The student groups taking the tests included fourth, sixth and eighth graders, high school and college students, graduate students in psychology, medical school students, and applicants to graduate business management programmes. The topics have included heart and circulatory anatomy and physiology (the same prompt at student levels from sixth grade to medical school in various studies), neural conduction, Pavlovian and operant conditioning, aphasia, attachment in children, Freudian concepts, history of the great depression, history of the Panama Canal, ancient

TABLE 1. Scoring of two Graduate Management Aptitude Test (GMAT) topics, an argument and opinion essays, the narrative essay described in the text, and several classroom essays

| *Standardised* | N | Reader 1 to Reader 2 | IEA-Single Reader |
|---|---|---|---|
| gmat1.train | 403 | 0.87 | 0.88 |
| gmat1.test | 292 | 0.86 | 0.84 |
| gmat2.train | 383 | 0.85 | 0.83 |
| gmat2.test | 285 | 0.88 | 0.85 |
| narrative.train | 500 | 0.87 | 0.86 |
| narrative.test | 400 | 0.86 | 0.87 |
| *Classroom* | | | |
| great depression | 237 | 0.65 | 0.77 |
| heart | 188 | 0.83 | 0.77 |
| aphasia | 109 | 0.75 | 0.66 |
| attachment | 55 | 0.19 | 0.54 |
| operant | 109 | 0.67 | 0.69 |
| freud | 239 | 0.89 | 0.78 |
| rogers | 96 | 0.88 | 0.68 |
| All Essays | 3296 | 0.83 | 0.81 |
| Standardised | 2263 | 0.86 | 0.85 |
| Classroom | 1033 | 0.75 | 0.73 |

American civilisations, alternative energy sources, business and marketing problems, age-relevant opinion pieces, creative narrative composition tasks in which students were given scenarios or two-sentence beginnings of a story and asked to complete it, and others. Overall, IEA scores have correlated on average within two percentage points of the correlations between two human graders. Table I shows comparative results for 13 representative sets. In all cases, these correlations are between IEA scores and human scores for a large held-out sample of essays, that is, ones not used in training the system. (Note also that these data do not include many studies from more recent commercial applications whose results cannot be shared, and which are based on improved techniques and produce somewhat better IEA performance.)

The average correlation between two independent readers and the average correlations of IEA scores with each of the two separately are shown. IEA correlations with the means of the human graders—not shown—are, expectedly, somewhat higher, for example, 0.90 in the case of the large set of reserved test narratives. This is an important result, because it shows that when trained on data from multiple readers, an IEA score used as a second reader on new essays will agree better with one human than another human would.

Taken together, the results indicate that overall IEA to human reliabilities were the same as human to human reliabilities within probable measurement error. The three major scoring components were all well correlated with human scores: on average r = 83, 0.68 and 0.66 respectively for content, style and mechanics scores [2]. Because of their redundancy, all variables together gave only marginally better prediction than the content component alone, and the empirically optimum weights for style and mechanics scores were much lower than that of the content component, on average 13% and 11%, versus 76%, when combined by empirical multiple linear regression. The content component alone accounted for three times as much variance for professional testing service exams, among which the narrative and argument essays did not require common factual content. Content accounted for four times as much variance for the more factually focused classroom-type tests.

## Other Empirical Validations of IEA Accuracy

In several ways IEA has appeared as valid or more so than human essay scores:

1. LSA's scores for heart anatomy and function essays predicted short answer test scores, r = 0.76; essay scores by two professional readers, 0.72 and 0.81 (Wolfe *et al.*, 1998).
2. Heart essays were written both before and after reading an instructional document. The professional ETS consultant graders were instructed to grade for content, and for this analysis, only the IEA content score was used. Differences in mean scores were measured in units of pooled within-group standard deviations. The difference by single human grader scores averaged 1.06 s.d., those by IEA content scores, 1.58 s.d., n = 94, by t-test $p < .003$. Note that the IEA model was not built explicitly to measure differences between the essays written before and after reading; but to predict human scores. Thus it was IEA's predicted human content grades that were more discriminating than the human scores themselves.
3. For 900 narrative essays written by primary school students in Grades 4, 6 and 8 we used discriminant function analysis to predict grade levels on the basis of (i) the average of two human reader grades, and (ii) IEA predicted grades. Leave-one-out cross-validation values (a statistical technique in which the predicted score is never part of the training set) gave 66% correct classification using human grades, and 74% using IEA. Assuming binomial distribution of these values, the IEA scores performed the classification five standard deviations better than the human grades. A likely reason is greater stability of estimate resulting from the use of multiple comparison essays. Discriminant function on age-in-days of students classifies the same cases 98.5% correctly. The correlation of IEA with age was .72, that of mean human scores .69.
4. In a large undergraduate class, ten-minute student essays on neural conduction were graded by undergraduate teaching assistants, graduate teaching assistants, and/or the professor, all blind to the IEA grades (using an early, content-only version of the LSA-based grader). The machine grader was trained on the mean of two teaching assistant grades. Correlations between machine and human were 0.69 for undergraduate teaching assistants, 0.78 for teaching assistants, and 0.80

for the professor. The difference between undergraduate and graduate students, based on over 300 scores in each case, is significant, $p < .01$.

## Discussion of Validity Studies

The first approach, correlation with another type of test, needs little discussion. The value of the procedure depends on being able to interpret criterion measures and their desirable degree of relation to essay content. In the above case, the equal prediction of a short answer test by humans and IEA adds confidence that essays scored either way were related to the knowledge in question.

The second approach, comparing human and machine scoring methods on their ability to discriminate between groups that should differ on the measured trait, resembles the calibration of IQ measures against school grade. In the reported cases, it showed that the IEA measure was a much more sensitive differentiator of learning experience. The potential limitation in such an approach is, of course, the need to assume that the test is measuring differences in abilities of interest rather than spurious associated variables. That the IEA score is based on and calibrated against human scores gives considerable assurance in this respect.

The third approach, comparing relative scoring success of the automatic grader to less and more highly qualified human judges offers a different perspective on validity. The approach would be more efficient with a more completely balanced design than was implemented here.

## The Range of Prompts and Topics Suitable for Automatic Scoring

There has been some worry that automatic essay grading would penalise or undervalue creativity. The IEA was originally designed for factual knowledge, for example, essays on anatomy, accounts of historical events, or descriptions of how to repair a jet engine, in which high creativity is usually not desired. However, in its application to ‘opinion’ essays on the GMAT and to narrative fiction essays, there was ample opportunity for exercising creativity, yet IEA was as reliable as the professional readers. For the $> 800$ middle-school creative narratives administered and independently scored by a professional testing organisation, blind IEA scores correlated $r = 0.90$ with averaged highly trained expert grader scores. How could this be? One hypothesis is that the constant setting for the story permitted only a limited variety of themes, plots, and characters, ones that drew upon the cultural experience and literary knowledge of the schoolchildren. Moreover, LSA can capture the similarity of stories that differ only in irrelevant details. For example, IEA might represent a humanly scored story about ‘two boys in search of rocks’ as one of the $k$ most similar to a comparison story about ‘some girls looking for roses’ (LSA similarity $= 0.63$) $p < .0001$. And, as mentioned earlier, IEA compares every essay to every other, something the humans did not and could not do for the set of 800!

Nevertheless, especially with essays for which there are fewer pre-scored papers, it appears prudent to construct prompts so as to delimit the variety of different content, for example to ask 'What changes would you like in the courses at your high school and why?, rather than 'What changes in the world … ?'

## Discussion and Summary

Despite pioneering work dating back several decades (Page, 1966), automated technology for analysis and scoring of open-ended written work is still in its infancy. This is largely because the computing power and software sophistication required to do the job well has only recently been available. However, in the last five years, with the working demonstration systems for IEA available on the Web in 1996, Intellimetric service offerings in 1998, and ETS's e-rater applications to GMAT scoring in 1999, it has become a reality.

It has been a considerable surprise not only to the public and educators, but even to many researchers that automated scoring could so easily, and by so many apparently different routes, produce reliabilities equivalent to that of humans. This raises a suspicion that a ceiling effect may be at work, that human graders are so unreliable that they can be matched by measuring almost anything that distinguishes good from bad essays. This hypothesis seems especially applicable to evaluation studies in which reliabilities have been below about 0.80. However, there have now been reports of reliabilities of around 0.9 by both machine and expert humans on the same sets of essays. This seems especially remarkable in that the cases involved were essays ostensibly subject to great variation in content and creativity, e.g. the GMAT argument and opinion essays scored by e-rater and IEA, and the free-form narrative stories scored by IEA. To continue being sceptical, this raises an alternative hypothesis, that the job is just too easy, that differences between essays are very large and easily detected (although the difference in reliability for undergraduate and graduate student teaching assistants described above argues otherwise). One way to resolve the joint implication of these observations is to suppose that: (a) qualitative differences between essays are highly redundant—good students do everything right, poorer students do everything less well; (b) the amount of qualitative difference among essays varies strongly with the kind of essay and writer population; and (c) there is a fairly constant amount of inter-judge disagreement, perhaps largely due to legitimate differences in opinion. Thus, if there are small real differences among the essays—for putative example in narrow factual essays—relatively constant inter-judge variability will dominate, while if the essays vary greatly—for putative example in open-ended narrative—the reliabilities of both humans and machine will be much higher. This is a rather optimistic interpretation in that it suggests that the underlying human disagreement component may not be as large as one would have thought. The evidence just cited, for example, implies that when essay quality varies sufficiently, human error can be less than 20% of the variance. This suggests a principle that essay prompts be designed to induce wide variations in quality. Of course, this is not psychometric news!

The good news, then, is that true differences in essay quality are apparently easy to detect. The bad news, as they say, is the same. The fact that all these disparate methods produce similar results at least raises the suspicion that one could get acceptable reliability results with the use of variables that are not conceptually valid. For example, because of high redundancy, it is not unlikely that counting the number of commas and semicolons in an essay would yield reasonably high correlations with human grades. What would be the matter with that? Many things. It would be open to easy coaching and forgery. It would have low face validity, attract legitimate ridicule, and undermine public acceptance. Worst, it would pull the evolution of question and prompt design, and student preparation, in erratic and possibly irrelevant directions. To a lesser extent, using any variable in automatic scoring that has low authenticity as a factor that educators want to promote holds such danger. Combining many variables, each with a different focus, is one way to ameliorate this problem, but will only suffice if the differences among the variables are due to more than measurement error. To date, we have seen no internal or external analyses of measured variables that adequately address this issue.

However, in the meantime, we think there is immediate worthy employment for the existing technologies. Here are just a few examples. Used as second or third opinions on high-stakes exams, they can increase consistency and decrease bias in scoring. Used for practice for high-stakes essay tests, their much lower cost, wider availability, and capability to provide instant scores can level the playing field. Used as components of interactive knowledge and writing tutorial systems, they can vastly increase the amount of useful practice in reading, learning, thinking and writing that students can engage in.

In sum, we think that the present is bright, the future, with the right effort, much more so.

## Acknowledgements

The basic science underlying IEA originated in research in the Institute of Cognitive Science at the University of Colorado at Boulder. Research on educational applications of IEA has been supported by the National Science Foundation, the ARPA Computer Aided Education and Training Initiative, the McDonnell Foundation Cognitive Studies for Educational Practice program, the Army Research Institute, the Air Force Office of Scientific Research, the Office of Naval Research and NASA. We gratefully acknowledge the collaboration of Intelligent Automation Inc., the Ontario Institute for Studies of Education, Robert Sternberg, Yale, and the members of the SALSA research groups at the University of Colorado and New Mexico State University, including especially, but not limited to, Walter Kintsch, Eileen Kintsch, Bob Rehder, and David Steinhart.

## NOTES

[1] A well known psychological 'law' (Miller, 1956) states that a human judgement of absolute values on any single dimension is limited to about three bits of accuracy, thus less than eight ordinal categories. Machine grades for essays have no such limit.

[2] These data are taken from all IEA applications in which there was more than one reader during the past five years of its development. They may underestimate current accuracy.

## REFERENCES

BERRY, M. W. (1992) Large scale singular value computations, *International Journal of Supercomputer Applications*, 6 (1), pp. 13–49.

BERRY, M. W., DUMAIS, S. T. & O'BRIEN, G. W. (1995) Using linear algebra for intelligent information retrieval, *SIAM: Review*, 37 (4), pp. 573–595.

BURSTEIN, J. (2003) The *e-rater*® scoring engine: Automated essay scoring with natural language processing, in: M. D. SHERMIS & J. BURSTEIN (Eds) *Automated essay scoring: a cross-disciplinary perspective* (Hillsdale, NJ, Lawrence Erlbaum Associates, Inc.), pp. 113–122.

CLAUSER, B. E., ROSS, L. P., CLYMAN, S. G., ROSE, K. M., MARGOLIS, M. J., NUNGESTER, R. J. PIEMME, T. E., CHANG, L., EL-BAYOUMI, G., MALAKOFF, G. L. & PINCETL, P. S. (1997) Development of a scoring algorithm to replace expert rating for scoring a complex performance-based assessment, *Applied Measurement in Education*, 10, pp. 345–358.

CLYMAN, S. G., MELNICK, D. E. & CLAUSER, B. E. (1995) Computer based case simulations, in: E. L. MANCALL & P. G. BASHOOK (Eds) *Assessing Clinical Reasoning: the oral examination and alternative methods* (Evanston, IL, American Board of Medical Specialties).

DEERWESTER, S., DUMAIS, S. T., FURNAS, G. W., LANDAUER, T. K. & HARSHMAN, R. (1990) Indexing by latent semantic analysis, *Journal of the American Society for Information Science,* 41 (6), pp. 391–407.

DOLIN, R., PIERRE, J., BUTLER, M. & AVEDON, R. (1999) Practical evaluation of IR within automated classification systems, *CIKM 99* (Kansas City, MO, ACM).

ECKART, C. & YOUNG, G. (1936) The approximation of one matrix by another of lower rank. *Psychometrica*, 1, pp. 211–218.

FOLTZ, P. W. (1996) Latent Semantic Analysis for text-based research, *Behavior Research Methods, Instruments and Computers,* 28 (2), pp. 197–202.

FOLTZ, P. W., KINTSCH, W. & LANDAUER, T. K. (1998) The measurement of textual coherence with Latent Semantic Analysis, *Discourse Processes*, 25, pp. 285–308.

GOLUB, G. & VAN LOAN, C. (1989) *Matrix Computations* (Baltimore, MD, Johns Hopkins).

KINTSCH, E., STEINHART, D., STAHL, G., MATTHEWS, C., LAMB, R. & THE LSA RESEARCH GROUP (2000) Developing summarization skills through the use of LSA-based feedback, *Interactive Learning Environments*, 8 (2), pp. 87–109.

LAHAM, D. (1997) Latent Semantic Analysis approaches to categorization, in: M. G. SHAFTO & P. LANGLEY (Eds) *Proceedings of the 19th Annual Meeting of the Cognitive Science Society* (Mahwah, NJ, Erlbaum).

LAHAM, D. (2000) Automated content assessment of text using Latent Semantic Analysis to simulate human cognition. Ph.D. Dissertation, University of Colorado, Boulder.

LAHAM, D., BENNETT, W. JR. & LANDAUER, T. K. (2000) An LSA-based software tool for matching jobs, people and instruction, *Interactive Learning Environments*, 8, pp. 1–15.

LANDAUER, T. K. & DUMAIS, S. T. (1997) A solution to Plato's problem: the Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge, *Psychological Review*, 104, pp. 211–240.

LANDAUER, T. K., LAHAM, D. & FOLTZ, P. W. (2003) Automated scoring and annotation of essays with the Intelligent Essay Assessor, in: M. SHERMIS & J. BURSTEIN (Eds) *Automated essay scoring: a cross-disciplinary approach* (Mahwah, NJ, Lawrence Erlbaum Associates, Inc.), pp. 87–112.

MILLER, G. A. (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information, *Psychological Review*, 63, pp. 8–97.

PAGE, E. B. (1966) The imminence of grading essays by computer, *Phi Delta Kappan*, 48, pp. 238–243.

PAGE, E. B. (1994) Computer grading of student prose, using modern concepts and software, *Journal of Experimental Education*, 62, pp. 127–142.

REHDER, B., SCHREINER, M. E., WOLFE, B. W., LAHAM, D., LANDAUER, T. K. & KINTSCH, W. (1998) Using Latent Semantic Analysis to assess knowledge: some technical considerations, *Discourse Processes*, 25, pp. 337–354.

WOLFE, M. B., SCHREINER, M. E., REHDER, B., LAHAM, D., FOLTZ, P. W., KINTSCH, W. & LANDAUER, T. K. (1998) Learning from text: matching readers and text by Latent Semantic Analysis, *Discourse Processes*, 25, pp. 309–336.